# ANÁLISIS DE LAS BASES DE DATOS NOSQL COMO ALTERNATIVA A LAS BASES DE DATOS SQL

CARLOS ANDRÉS LÓPEZ PEÑA

**Trabajo de grado para optar al título de ingeniero informático**

**Santiago Villegas Giraldo**


**ESCUELA DE INGENIERÍA DE ANTIOQUIA**
**INGENIERÍA INFORMÁTICA**
**ENVIGADO**
**2012**

# AGRADECIMIENTOS

Agradezco a mi familia por el apoyo brindado a lo largo de toda la carrera que me ha permitido llegar hasta la culminación de la profesión de ingeniería informática, quiero agradecer también a todos los profesores tan capaces y excelentes que tiene la universidad y al director del trabajo de grado y las personas de las empresas encuestadas que me guiaron y colaboraron durante la realización del trabajo de grado.

# CONTENIDO

pág.

# LISTA DE TABLAS

pág.

# LISTA DE FIGURAS

pág.

# LISTA DE ANEXOS

pág.

# GLOSARIO

ACID: acrónimo de Atomicity, Consistency, Isolation, Durability. Modelo que garantiza la atomicidad, consistencia, aislamiento y durabilidad de la base de datos. Esto es que todas las operaciones se deben realizar completamente o de lo contrario se regresa el sistema al estado antes del cambio; los datos a modificar sean los correctos; todas las operaciones sean independientes entre sí; y los cambios realizados exitosamente perduren.

ANS: acrónimo de Acuerdo de Nivel de Servicio (SLA en inglés). Contrato con el cliente sobre algunas características del sistema como disponibilidad, tiempos de respuesta, rendimiento, etc.

API: traducido como Interfaz de Programación de Aplicaciones, son librerías para ser utilizadas por otras aplicaciones o servicios que permiten el uso de métodos, funciones, procedimientos, etc.

AWS (Amazon Web Services): Conjunto completo de servicios de infraestructuras y aplicaciones que permiten ejecutar prácticamente todo en la nube. (Amazon.com Inc.)

Back up: Respaldo de información.

BASE: acrónimo de Basically Available, Soft state, Eventual consistency. Modelo alternativo a ACID que es básicamente disponible, de estado ligero y eventualmente consistente, es decir que no asegura la disponibilidad, el estado del sistema puede cambiar eventualmente incluso sin modificaciones y finalmente, el sistema llegará a un estado de consistencia con el tiempo mientras no reciba más ingresos.

BI: Business Intelligence, traducido como Inteligencia de Negocios, es un término que incluye aplicaciones, infraestructura, herramientas y las mejores prácticas que permiten el acceso y análisis de la información para mejorar y optimizar decisiones y rendimientos. (Gartner)

Column Families: modelo avanzado de la estructura key-value. Similar a una tabla en un modelo de bases de datos relacional (RDBMS), contiene filas y columnas. Cada fila es identificada únicamente por una clave de fila; cada fila tiene múltiples columnas, y cada una de estas tiene un nombre, valor y estampa de tiempo. (Apache Cassandra 0.7 Documentation)

Confidencialidad: la información no se pone a disposición ni se revela a individuos, entidades o procesos no autorizados. (López Neira & Ruiz Spohr)

DBA: acrónimo de DataBase Administrator (Administrador de Bases de Datos). Persona encargada de realizar el mantenimiento y aseguramiento de los datos en un sistema de gestión.

Disponibilidad: acceso y utilización de la información y los sistemas de tratamiento de la misma por parte de los individuos, entidades o procesos autorizados cuando lo requieran. (López Neira & Ruiz Spohr)

Document database: almacenamiento de datos como documentos, similar a los objetos en JavaScript.

EIDE: acrónimo de Enhanced Integrated Drive Electronics (Unidad Electronica Integrada Mejorada). También conocidos como PATA (Parallerl Advanced Technology Attachment), la etiqueta en estas unidades corresponde a la alternativa de interfaz en la que involucra conectar el tablero de la UCP y la unidad. Formatear este tipo de unidades puede tomar un largo tiempo ya que su velocidad es reducida. (Shukla, 2011)

EMR: Amazon Elastic MapReduce, implementa Apache Hadoop y Apache Hive por medio de MapReduce para controlar la complejidad de las aplicaciones además se integra con otros servicios de AWS.

Escalabilidad: capacidad de mejorar recursos para ofrecer una mejora en la capacidad de servicio. (MSDN)

ETL: traducido como Extracción, Transformación y Carga, es un proceso que permite manipular para facilitar el uso en diferentes sistemas.

Ext3: Extended 3, o tercer sistema de archivos extendido, versión mejorada de ext2. Es un sistema de archivos que trabaja con registros por diario y utiliza un árbol binario balanceado

Ext4: Extended 4, o cuarto sistema de archivos extendido, versión mejorada de ext3. Es un sistema de archivos con registro que soporta volúmenes de almacenamiento de hasta 1024 PiB, usa menos recursos del sistema y mejora la lectura, escritura y verificación de los archivos. (Barrios Dueñas)

Framework: traducido como marco de trabajo, espacio designado para que los usuarios interactúen con el sistema proveyendo ventajas de usabilidad y funcionalidad.

Graph database: almacena datos en grafos, la estuctura más genérica de las estucturas de datos, capaz de representar cualquier dato de forma altamente accesible. Se leen los grafos siguiendo las flechas direccionales alrededor del diagrama para formar sentencias. (Neo Technology). Un grafo es una manera avanzada de ver el almacenamiento key-value, ya que este primero se forma cuando los valores entre ellos mismos están interconectados.

GUI: Graphic User Interface, traducido como interfaz gráfica de usuario, es una herramienta o marco de trabajo visual para la interacción con el usuario.

Integridad: mantenimiento de la exactitud y completitud de la información y sus métodos de proceso. (López Neira & Ruiz Spohr)

IRC: Internet Relay Chat. Sistema para chatear de tipo cliente-servidor.

JSON: JavaScript Object Notation. Es un formato para almacenar e intercambiar datos, independiente del lenguaje de programación que se esté utilizando. (JSON Tutorial)

JRE: Java Runtime Enviroment, traducido como ambiente de ejecución de Java, es necesario que se encuentre instalado en las máquinas para correr aplicaciones basadas en Java.

Key Value database: modelo que se puede considerar como 2 columnas, una con el valor de la "clave" –conocido como id o identificador- y otra con el valor "real" al que referencia la llave.

Little-endian: formato en el sistema operativo en el que el byte de menor peso se almacena en la dirección más baja de memoria y el byte de mayor peso en la más alta.

Madurez: experiencia y conocimiento que tiene una empresa.

Mantenibilidad: según la IEEE (Institute of Electrical and Electronics Engineers, 1990), se define como "facilidad con la que un sistema de software o componente puede ser modificado para corregir fallos, mejorar el rendimiento u otros atributos o adaptarse a cambios en el entorno".

MapReduce: Consta de 2 partes "map" y "reduce". Esta divide los flujos de trabajo (map) para procesarlos en paralelo y recombinar los datos procesados en la solución final (reduce). (Amazon.com Inc.)

Normalización: cuando se asocia a las bases de datos, es un término que se utiliza usualmente durante el diseño de software por medio del modelo entidad relación. Existen varios niveles o formas de la normalización para organizar los datos, permitir flexibilidad y evitar redundancias.

NoSQL: definido como Not only SQL, es un tipo de sistema de gestión de bases de datos que pretende ser la siguiente generación sobre estas tecnologías.

Replicación: copiar los datos de una ubicación a otras ubicaciones para evitar perder información.

RUP: Rational Unified Process, traducido como Proceso Racional Unificado, es un proceso iterativo e incremental de buenas prácticas para el ciclo de vida del desarrollo de software, desde la planeación hasta despliegue.

SATA o Serial ATA: acrónimo de Serial Advanced Technology Attachments (Componente de Tecnología Avancada Serial). Era una tecnología de bus primordialemten diseñada para transferir datos hacia y desde el disco duro. Contiene 2 conectores separados, uno para datos y otro para la energía, aunque puede tener un tercero para las conexiones de energía para PATA. (Serial ATA Connector Pinout)

SCSI: acrónimo de Small Computer System Interface (Interfaz de Sistema para Pequeños Computadores). Su funcionalidad es similar a los EIDEs ya que utilizan discos magnéticos rotatorios para escribir y almacenar datos en este. La mayor diferencia entre SCSI, EIDE y SATA es que el tipo de disco SCSI gira más rápido por lo que procesa y almacena los datos más rápidamente, sin embargo es más propenso a dañarse debido a la velocidad de rotación. (Shukla, 2011)

SSD: acrónimo de Solid State Drive (Unidad de Estado Sólido). Compuesto de parte estacionarias opuesto a otras unidades que generalmente incluyen discos magnéticos para rotar mientras respaldan los datos. Están equipados con semiconductores que realizan la tarea de rotar discos magnéticos para apilar los datos almacenados. Al no poseer partes móviles son menos susceptibles de dañarse, además la velocidad de procesamiento es más rápida que otras unidades. (Shukla, 2011)

Tabla: en bases de datos, es un elemento en la cual se organizan y almacenan los datos en celdas, los valores de estas celdas están asociados a una fila y a una columna.

UCP: Unidad Central de Procesamiento. Parte de un sistema informático que interpreta y procesa los datos generados por los programas informáticos.

Usabilidad: según la IEEE (Institute of Electrical and Electronics Engineers, 1990), se define como “facilidad con la que un usuario puede aprender a operar, producir entradas e interpretar las salidas de un sistema o componente”.

Versión: cuando se le atribuye al software, se interpreta como el estado o etapa en la que se encuentra una aplicación o servicio prestado.

ZFS: sistema de archivos que cambia la forma como los sistemas de archivos son administrados, es robusto, escalable y fácil de administrar.

# RESUMEN

A medida que pasa el tiempo, la demanda de almacenamiento y velocidad de operaciones en las bases de datos se ha incrementado, esto se debe a que más y más empresas tienen más aplicaciones, más clientes y quieren dar un mejor servicio.

En algunas empresas la demanda de los servicios de bases de datos es tan elevada que no se pueden permitir hacer esperar al usuario, y como solución hasta el momento –en algunas empresas- es optimizar el sistema de información, esto es tanto a nivel lógico como físico, desde verificar y optimizar las consultas internamente hasta adquirir servidores con mejores especificaciones que permitan responder rápidamente a la demanda. Como otra solución, surgió una nueva tecnología denominada NoSQL, que se desprende del estándar por el que estaban regidos los anteriores motores de bases de datos, logrando mejores resultados en la realización de operaciones al motor como escritura y lectura de datos.

Este trabajo de grado pretende proponer una alternativa NoSQL a los sistemas de gestión de bases de datos que utilizan las empresas en Medellín, de forma que puedan ser competitivas y hagan uso de las tecnologías emergentes que abordan las necesidades de hoy en día.

Para lograr esta propuesta, se realizan varias encuestas a personas que estén relacionadas con la adquisición, operación y mantenimiento de las bases datos en organizaciones con alto grado de madurez y que se encuentren en la ciudad de Medellín, de tal forma que permitan conocer las necesidades y requisitos organizacionales para posteriormente optar por una opción de bases de datos NoSQL.

Palabras clave: Bases de datos, NoSQL

# ABSTRACT

As the time passes, the storage demand and speed of transactions in the databases has increased, this is because more and more enterprises have more applications, more clients and they want to give a better service.

But in some enterprises the service demand of databases is so high that they can´t allow to make wait the user, and up until now –in some enterprises- have optimized the information system, this is from logic layer to physic layer, from verity and optimize queries to acquire servers with better requirements allowing agile responses to the demand. As another variant of solution, a new technology appeared named NoSQL, that doesn't follow the standard the older databases engines had, achieving better transactions results on the engine like data writing and reading.

In this degree work, I pretend to propose a NoSQL alternative to the database management system used by enterprises in Medellin; this way, they can be competitive and use emerging technologies.

To achieve this proposal, some surveys were carried out to persons related to the database acquisition, operation and maintenance in high level maturity organizations located in the city of Medellín, in order to realize their needs and requirements and subsequently find at least one NoSQL database option.

Key words: Databases, NoSQL

# INTRODUCCIÓN

Actualmente las organizaciones que manejan bases de datos operan con gran cantidad de datos, esto se debe a la cantidad de usuarios, aplicaciones y necesidades internas o externas. Tanto es el crecimiento de estas tecnologías que no es mantenible ni escalable continuar en algunos casos con los mismos motores de bases de datos ya que dificulta la administración y mantenimiento de la información.

Hasta ahora se pueden identificar a grandes rasgos dos tecnologías de gestión para el almacenamiento de datos: Las basadas en SQL y las que no siguen este estándar, denominadas NoSQL; las primeras son las mayormente conocidas como MySQL y Oracle de la compañía Oracle, SQL Server de Microsoft, PostgreSQL de PostgreSQL Global Developer Group.

Las bases de datos NoSQL surgieron como consecuencia de los rendimientos de los motores basados en SQL, ya que no cumplían las expectativas de tiempo de respuesta. Esta siguiente generación de sistemas de gestión de bases de datos contiene unas características particulares, ya que es no relacional, distribuida, de código abierto y horizontalmente escalable –estas características se explicarán en detalle más adelante-, de esta forma se optimiza el tiempo de consulta sobre las bases de datos y soportan una mayor cantidad de datos.

Adicionalmente esta generación presenta divisiones: se pueden encontrar bases de datos de tipo *key-value, column families*, almacenamiento de documentos, gráficas, entre otras más, pero se trabajará principalmente sobre estas ya que son las más comunes de encontrar.

Para el trabajo de grado se seleccionaron las más relevantes, correspondientes a: Cassandra, MongoDB y Neo4j. Sobre estas se investigará sus características de acuerdo a varios factores como la mantenibilidad, características de la máquina, instalación e implementación, usabilidad, versión, soporte, madurez, documentación y características adicionales.

Con esta información presente, y la realización de encuestas a diferentes empresas en Medellín relacionadas con las características que requieren a la hora de adoptar los sistemas de gestión de bases de datos, se procederá a proponer una alternativa de base de datos NoSQL.

# 1. PRELIMINARES

En este capítulo se retoma la información presentada en el anteproyecto y se mejora con el fin de hacer una presentación clara de lo que se aprobó y lo que se realizó.

## 1.1 PLANTEAMIENTO DEL PROBLEMA

La cantidad de información que se almacena en las bases de datos de grandes compañías, así como la cantidad de operaciones (lectura, escritura, modificación y eliminación) por parte de los usuarios es enorme, por lo que las compañías se han quedado cortas para responder rápidamente a las necesidades de sus usuarios. Además no es mantenible ni escalable soportar estas bases de datos con tal presión, ya que dificulta la administración por parte de los DBA.

Como respuesta, se originó el concepto de NoSQL, que pretende ser la siguiente generación de tecnologías de sistemas de gestión de bases de datos.

Se requiere entonces explorar algunas bases de datos NoSQL representativas, como Cassandra, MongoDB y Neo4j, como alternativa a los sistemas de gestión de bases de datos, para hacer frente a la cantidad de datos y número de peticiones requeridas por los clientes (tráfico de datos) en empresas en Medellín.

De acuerdo con los precios, características y requerimientos encontrados se puede dar la recomendación de implementar alguno de estos sistemas de gestión de bases de datos NoSQL.

## 1.2 OBJETIVOS DEL PROYECTO

### 1.2.1 Objetivo General

Analizar características existentes en una base de datos NoSQL que permitirían a una empresa utilizarlas en vez de una base de datos SQL.

### 1.2.2 Objetivos Específicos

- Identificar y documentar las características que buscan las empresas a la hora de adoptar los sistemas de gestión de bases de datos.
- Verificar el cumplimiento de las características obtenidas en el objetivo específico anterior con los sistemas de gestión de bases de datos NoSQL.
- Proponer alternativas que cumplan las expectativas de las empresas con base en lo registrado con el cumplimiento de características en el objetivo específico anterior.

## 1.3 MARCO DE REFERENCIA

Para entender un poco NoSQL expliquemos primero la figura 1 obtenida a través de Google Trends, en la que se visualizan 2 campos, el primero, Search Volume index muestra la cantidad de búsquedas realizadas por personas, y el segundo, News reference volume, muestra la cantidad de veces que ha aparecido en el tópico de historias de Google News; las letras A y B hacen referencia a búsquedas artículos en internet que han tenido gran rating de visitas.

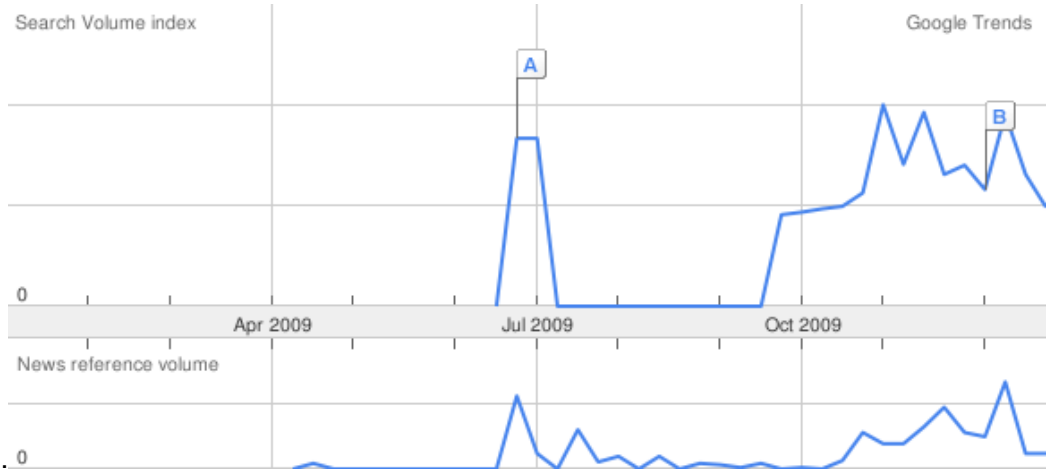


**Figura 1. Tendencias NoSQL**

Como se puede ver, NoSQL comenzó a tener popularidad a mediados de abril de 2009, por lo que es relativamente un concepto nuevo y la mayoría de la información que se puede obtener se basa en artículos, noticias y trabajos en formato virtual.

NoSQL puede ser definido como “Not only SQL” y es un sistema de gestión de bases de datos que pretende ser la siguiente generación de tecnologías de estas que es no relacional, distribuida, de código abierto, horizontalmente escalable y más rápida, ya que no implementa las propiedades ACID para asegurar la confiabilidad de las transacciones sobre las bases de datos. NoSQL fue definida inicialmente para modernizar las bases de datos en la nube pero se aplica en general, ya que por la cantidad de información que se almacena en ellas, pueden no responder tan rápidamente como se espera en horas pico o de alto tráfico de datos incumpliendo los ANS.

Entre las características que posee NoSQL se encuentra que no presentan esquemas, tienen fácil soporte de replicación, API simple, eventualmente consistente (conocido como BASE, y contrario al concepto de ACID) y contienen enormes cantidades de datos. (Información NoSQL)

Retomando lo que nos indica la Figura 1, al ser un concepto nuevo, se está incursionando en esta modalidad y se puede ver como Amazon y Oracle han creado bases de datos NoSQL; Twitter, Netflix, Facebook, Cisco, entre otros, utilizan estos tipos de motores; CouchDB y SQLite crearon un lenguaje de consulta llamado UnQL (Jackson, 2011). Hasta el momento existen muchos motores ya que son de código abierto, por lo que se

tendría que revisar cuales son los que presentan mejores funcionalidades, y reafirmando la conclusión de Orend, se tiene que analizar varias alternativas para determinar cuál es la que satisface mejor una necesidad.

Para dar una mayor calidad y pofundidad de las propiedades NoSQL mencionadas al principio de este trabajo se realizará una comparación frente a su término opuesto:

- No relacional vs. relacional: las bases de datos no relacionales son listas de datos almacenados en una sola tabla sin definir relaciones entre los registros. Por otro lado las relacionales reparten los datos en varias tablas más pequeñas eliminando datos duplicados y asegurando consistencia y estableciendo restricciones y relaciones con otras tablas por medio de claves primarias y foráneas; esto genera que se ocupe menos espacio ya que no tiene redundancias y hasta cierto punto es conveniente esta repartición ya que de otro modo, si se realiza un SELECT multiple en la no relacional se tendría que recorrer la única tabla varias veces para devolver todos la información debido a duplicidad en ciertos datos.

- Distribuida vs. centralizada: tener las bases de datos almacenadas en varios nodos para poder distribuir la carga en estos se denomina distribución (o descentralización), es decir que el almacenamiento se realiza en múltiples partes en una localización física, en una red interconectada o a través de internet como es el caso de cloud databases, sin compartir memoria o discos entre los nodos; en cambio en una base de datos centralizada se tiene un bus común a través de todos los nodos que comparte memoria, ésta se encuentra en una sola ubicación lo que facilita la administración, pero no la escalabilidad ya que genera cuellos de botella cuando residen muchas aplicaciones o usuarios.

- Código abierto vs. cerrado: como su nombre lo indica, abierto permite la visualización del código fuente por parte de los usuarios para su modificación creando grupos colaboradores para mejorar el código y compartirlo con las demás personas y puede ser distribuido gratuitamente o pagado.  no puede ser cobrado y la licencia debe ser libre; el código cerrado (o propietario) puede ser cobrado pero no permite el acceso al código fuente. Como consecuencia del código abierto en NoSQL, existe una gran oferta de propuestas en el mercado.

- Ejemplos de bases de datos propietarias: BigTable de Google, DynamoDB y SimpleDB de Amazon.

- Ejemplos de bases de datos libres: Cassandra de Facebook, CouchDB de Apache, Redis, Neo4j, MongoDB.

- Horizontalmente escalable vs. verticalmente: escalar horizontalmente significa obtener más nodos para un sistema, como adquirir un nuevo computador para una aplicación distribuida u comprar más servidores Web logrando crear cluster. En cambio, escalar verticalmente significa incrementar el número de recursos a un único nodo del sistema como adquirir mas CPUs o memoria para un computador o servidor. (Escalabilidad, 2012)

### 1.3.1 Bases de datos NoSQL

**Apache Cassandra**

Base de datos escrita en Java, de tipo Column Family, de código abierto por Facebook en 2008, diseñada por Avinash Lakshman (uno de los autores de Dynamo, de Amazon) y Prashant Malik (Ingeniero de Facebook). De varias maneras se puede pensar en Cassandra como Dynamo 2.0 o una unión de Dynamo y BigTable. Cassandra se encuentra en producción en Facebook, pero aún se encuentra bajo fuerte desarrollo.

Altamente escalable, eventualmente consistente, distribuida y almacenamiento estructurado key-value. Agrupa las tecnologías de sistemas distribuidos de Dynamo y el modelo de datos BigTable de Google. Como Dynamo, es eventualmente consistente. Como BigTable, provee modelo de datos basado en ColumnFamily más enriquecido que los sistemas comunes key-value. (Apache Cassandra, 2012)

Existen varias herramientas para la visualización y administración de los datos, la más destacada es OpsCenter que ofrece gestión y administración para los cluster, esta contiene una edición comunitaria y una empresarial que incluye características adicionales: alertas, balanceo automático de cargas, respaldos en vivo, entre otras.

Adicional a esta se pueden encontrar otras: Cassandra Cluster Admin, Cassandra Explorer y Helenos

Lenguaje de consulta: CQL (Cassandra Query Language)

Entre algunos usuarios se encuentran:

- Despegar: Usa un cluster de Cassandra para almacenar la sesión de los usuarios de su sitio de consulta de hotel, y también como cache persistente de itinierarios de viaje.
- Facebook: Tiene el mayor número conocido de clusters en operación con cerca de 150 nodos.
- Twitter: Almacenar y consultar la base de datos de lugares de interés; almacenar resultados de minería de datos sobre la base de usuarios; desarrollo interno y externo para análisis en tiempo real a gran escala. (King, 2010)
- eBay: Usado para soportar múltiples aplicaciones con anillos extendidos sobre varios data centers.

Sitio web oficial: cassandra.apache.org

***Características***:

Mantenibilidad: Al utilizar el lenguaje de consultas CQL, similar al estándar SQL facilita un poco el entendimiento e identificación para los desarrolladores que ya han utilizado otros motores.

Características de la máquina: Requiere la última versión estable de Java 1.6 JRE o una más actualizada. En cuanto a sistema operativo y versión no se encuentra explícitamente, por lo que depende de la instalación del ambiente de java.

Instalación/Implementación: En el sitio web oficial cassandra.apache.org se encuentra una sección dedicada a la descarga y otra a instalación y requerimientos necesarios.

Usabilidad: Comandos ingresados por medio de Cassandra CLI (interfaz de línea de comandos) o por medio de las herramientas gráficas y de gestión.

Versión: La última versión estable es la 1.1.5, lanzada el 10 de septiembre de 2012.

Soporte y solución de inquietudes: Contiene una sección llamada Wiki con información general y configuraciones; y otra nombrada FAQ (Frequently Asked Questions) con un listado de posibles preguntas. Adicionalmente contiene un vínculo IRC a irc.freenode.net - canales de chat gratuito- donde los desarrolladores y los miembros de la comunidad están disponibles para responder preguntas e inquietudes.

Madurez: Gran recorrido ya que fue inicialmente desarrollo interno de Facebook, y en 2008 salió a la luz como código abierto, se encuentra en versión estable. La primera versión para el público fue lanzada durante el segundo semestre de 2009.

Documentación: Como se mencionó en los ítems anteriores, contiene varios apartados que pueden guiar al usuario en diferentes actividades. Para el caso del lenguaje CQL, contiene un repositorio con las consultas soportadas, sus notaciones y respectivos ejemplos.

Ventajas: Orientada a ColumFamilies, tolerante a fallos, ya que replica los datos de forma automática a múltiples nodos; cuando un nodo falla pueden ser reemplazado sin ningún periodo de inactividad. Permite replicación a múltiples data centers; almacenamiento de los datos tipo ColumnFamily.

Limitaciones: El valor de una columna no debe ser mayor a 2GB, el máximo número de columnas por fila es de 2 billones, la llave y los nombres de las columnas deben ser menores a 64 KB.

**MongoDB**

Base de datos creada por la compañía 10gen, su nombre proviene de la palabra “hu**mongo**us” que se traduce como enorme.

Escrita en lenguaje C++, de código abierto, orientada a documentos, y pensada para ser escalable y de desarrollo ágil. Contrario a almacenar los datos en tablas y filas como se realizaría en una base de datos relacional, MongoDB almacena documentos similares a JSON con esquemas dinámicos. La finalidad es cerrar la brecha entre almacenamiento key-value (rápido y escalable) y las bases de datos relacionales (ricas en funcionalidades). (10gen, 2011)

MongoDB no incluye interfaz administrativa al estilo GUI, estos son realizados desde la misma consola mongo shell pero sí existen proyectos comunitarios aparte que han elaborado varias herramientas para la administración y visualización de datos. La mayoría son de distribución libre como Edda, Fang of Mongo, Umongo, etc., aunque también existen dos opciones comerciales de Nucleon Software: Database Master y Business Intelligence Studio –no sólo para MongoDB, sino también para bases de datos relacionales como Oracle, DB2, SQL Server, entre otros-.

Lenguaje de consulta: Mongo Query Language

Entre algunos usuarios se encuentran:

- Chartbeat: Almacenar datos analíticos historicos.
- MTV Networks: Administrar y distribuir contenido de los sitios web de MTV Networks.
- SourceForge: Almacenamiento de las páginas iniciales, de proyectos y de descargas.
- The New York Times: Aplicaciones hechas para la presentación de fotos.

Sitio web oficial: www.mongodb.org

***Características***:

Mantenibilidad: Por medio de la consola, se ingresan las sentencias que están basadas en JavaScript, por lo que facilita el entendimiento para aquellos que suelen usarlo. Consultas expresadas como objetos JSON a través del driver de sintaxis de Mongo C++, por lo que también se pueden generar consultas por medio de sentencias en este lenguaje.

Características de la máquina: Sistemas operativos OS X, Linux, Windows de 32 y 64 bits para todos los anteriores y Solaris de 64 bits

Instalación/Implementación: Sección MONGO DOCS, donde se puede encontrar información de la instalación, descargas y drivers.

Usabilidad: Comandos de consulta ingresados por medio de la consola por defecto, pero se pueden descargar o comprar herramientas gráficas y de análisis de otros grupos comunitarios o empresas de software.

Versión: La última versión estable es la 2.0.6, lanzada el 6 de Junio de 2012.

Soporte y solución de inquietudes: Sección MONGO DOCS CON, tutoriales, foros de ayuda en los grupos de Google y chat por medio del IRC irc.feenode.net/#mongodb. Sección FAQ para los desarrolladores, explicando conceptos de la base de datos. Como valor agregado se puede encontrar la sección TRY IT OUT para ensayar de forma online y gratuita la consola de comandos a modo de tutorial.

Madurez: A pesar de que es utilizado por algunas grandes empresas, no es muy reconocido. La primera versión para el público fue lanzada en el segundo semestre de 2009.

Documentación: Varias secciones de ayuda y soporte para los desarrolladores, sección especial orientada a los desarrolladores nombrada SQL to Mongo Mapping Chart –Esta última es la alternativa al lenguaje de consulta SQL- donde en términos generales, contiene 2 columnas, la primera cómo son las sentencias de consulta y la segunda columna correspondiente a las sentencias Mongo.

Ventajas: Orientada a documentos, comandos ingresados en la consola basados en JavaScript. Contiene también un driver en C++ que permite ingresar comandos en este lenguaje.

Limitaciones: Si se usa en distribuciones de 32 bits, los datos están limitados alrededor de 2GB. El servidor MongoDB debe ejecutarse en una CPU en little-endian. En lo posible usar Windows 2008 server en adelante, ya que tiene características nuevas que incrementan el rendimiento.

**Neo4j**

Base de datos gráfica (graph database) de alto rendimiento y desarrollada por Neo Technology. Esta almacena los datos en los nodos y relaciones de un grafo y opera con una estructura de red flexible y orientada a objetos en vez de tablas estáticas y rigurosas, pero a la vez obteniendo todos los beneficios de una base de datos transaccional.

A pesar de que es open source –correspondiente a la versión Comunitaria-, también cuenta con otras 2 licencias comerciales: Avanzada y Empresarial-, contando con mayores beneficios como soporte, monitoreo, recuperación de desastres, back up online y mayor disponibilidad.

Para las licencias comerciales, Neo Technology considera a Neo4j como un servicio, por lo que ofrece un acuerdo de servicio que incluye actualizaciones, corrección de bugs y asistencia para los proyectos.

Independiente de las versiones ofrecidas, Neo Technology ofrece otros servicios:

- Soporte de los productos: para despliegue de desarrollos.

- Servicios profesionales: servicios de asistencia como entrenamiento en los procesos de despliegue; integración con la arquitectura y lógica de la aplicación; y migración de datos, todo esto enfocado en el motor Neo4j

La interfaz gráfica que ofrece el propietario se llama Neoeclipse, un framework para la realización de los grafos

Lenguaje de consulta: Cypher Query Language

Sitio web oficial: neo4j.org

Entre algunos usuarios se encuentran:

- Adobe
- Cisco

***Características***:

Mantenibilidad: Lenguaje de consulta similar a SQL y adaptado al modelo key-value.

Características de la máquina:

| | **Sistema Operativo** | **CPU** | **Memoria** | **Disco** | **Sistema de archivos** | **Software** |
|---|---|---|---|---|---|---|
| **Mínimo** | Linux, Windows XP, Mac OS X | Intel 486 | 1GB | SCSI, EIDE | ext3 (o similar) | Java JRE 1.6 |
| **Requerido** | Linux y Windows: 32 y 64 bits | Intel Core i7 | 4GB | SSD con SATA | ext4, ZFS | Java JRE 1.6+ |

**Tabla 1. Requisitos Neo4j**

Los datos mostrados en la tabla 1 fueron obtetenidos del manual de Neo4j. (Neo Technology)

Instalación/Implementación: Manuel de Neo4j, con requerimientos, como los vistos en el ítem anterior,

Usabilidad: Herramienta visual llamada Neoeclipse, para la realización de los grafos.

Versión: La última versión estable es la 1.8, lanzada el 29 de septiembre de 2012.

Soporte y solución de inquietudes: Ítem Questions & Answers en la sección Resources, y enlace a StackOverflow –Sitio web de preguntas y respuestas-. Cuenta también con servicio de pago como soporte y asesoría, aunque las versiones comerciales ya lo incluyen en su compra.

Madurez: Implementaciones por compañías como Adobe y Cisco para potenciar nuevos servicios y lograr un mejor rendimiento. La primera versión para el público fue lanzada a comienzos de 2010.

Documentación: Sección llamada Docs, donde se puede encontrar información de las versiones paso a paso, desde qué es graph database, pasando por los componentes de un grafo, el lenguaje de consulta. Adicionalmente se encuentra la sección Resources, donde se encuentra el manual de Neo4j, ejemplos de grafos

Ventajas: Orientada a grafos, almacenamiento nativo optimizado para almacenar estructuras de grafos para un mayor rendimiento y escalabilidad, framework para la creación de los grafos, api orientada a objetos. Tiene varias opciones de licenciamiento de

software (Comunitaria, Avanzada y Empresarial) de acuerdo a las necesidades de la empresa.

Limitaciones: El soporte sólo está asociado a las versiones comerciales. Este tipo sistema NoSQL, según las implementaciones vistas, parece estar dirigido a soluciones donde se identifiquen nodos como redes sociales, mapas y localizaciones.

**DynamoDB**

Es un servicio de base de datos NoSQL abarcado por AWS (Amazon Web Services) y ofrecido por Amazon a comienzos del presente año (2012) de tipo key-value, que provee flexibilidad y alto rendimiento, disponibilidad y escalabilidad.

Esta base de datos está diseñada para abordar problemas de gestión, rendimiento, escalabilidad y fiabilidad de los datos ya que el cliente no realiza acciones de instalación, configuración, revisión, ni demás actividades de mantenimiento ni soporte del motor.

Para ayudar a gestionar las actividades propias de cada usuario como la infraestructura de la base de datos, los recursos informáticos y el almacenamiento, se cuenta con AWS Management Console, una interfaz web de tipo señalar y pulsar para diferentes acciones dentro del conjunto de servicios de Amazon Web Services.

Este motor, a diferencia de los anteriormente nombrados, este es de pago ya que cuenta con servicios provistos por Amazon, los datos se almacenan en sus servidores y es escalable en cuanto a que sólo se paga por lo que los usuarios usen, si se generan más solicitudes o concurrencia de usuarios, entonces Amazon destinará más servidores y continuará gestionando la seguridad de los datos y realizando operaciones de mantenimiento cuando sean necesarias.

De igual manera, Amazon provee la interfaz gráfica para usuario por medio de AWS Management Console para crear elementos en la base de datos y controlar recursos y medir el rendimiento sobre el motor

Sitio web oficial: http://aws.amazon.com/es/dynamodb/

***Características***:

Mantenibilidad: Altamente escalable y gestionado en cuanto a recursos necesarios por las aplicaciones que pueden acceder a la base de datos por parte de Amazon.

Características de la máquina: Navegadores web: Mozilla Firefox 12 y 13; Apple Safai 5; Google Chrome 19 y 20 y Windows Internet Explorer 8 y 9. Cabe resaltar que para hacer uso de AWS, no se admiten las versiones beta de los navegadores anteriores.

Instalación/Implementación: Tener cuenta de AWS.

Usabilidad: Uso de interfaz gráfica por medio de AWS Management Console.

Versión: Beta

Soporte y solución de inquietudes: La página web oficial contiene bajo la sección Support varios elementos de ayuda como foros, estado de los servicios de AWS, preguntas técnicas, entre otros elementos no sólo de este motor sino también de los demás servicios que ofrece AWS.

Madurez: Amazon como tal, contiene una amplia gama de soluciones por lo que es reconocido mundialmente y a pesar de que DynamoDB se encuentra en una versión beta para los usuarios, Amazon mismo utiliza esta misma base de datos y variantes para el procesamiento interno.

Documentación: Amazon tiene un manual de usuario llamado Amazon DynamoDB Developer Guide acerca de los conceptos e instrucciones para el uso de las operaciones en el motor.

Ventajas: No requiere administración sobre las bases de datos, ya que de esta parte se encarga el servicio de Amazon; por ejemplo: Adquisición de nuevo hardware, configuración general, replicación, aplicación de parches, configuración de particiones o clusterización.

Limitaciones: DynamoDB se encuentra en versión beta y no todos los navegadores de internet admitidos por AWS Management Console.

# 2. METODOLOGÍA

Para lograr el objetivo final de este trabajo de grado, el cual es identificar una alternativa viable NoSQL, es necesario un conocimiento suficiente de información tanto de bases de datos de este tipo como las SQL, así como los requisitos y necesidades de las empresas, para lograr este proceso se procedió como se muestra a continuación:

Inicialmente y aunque no hace parte de ningún objetivo específico, se realizó una búsqueda de bases de datos NoSQL y se seleccionaron 4 motores –pese a que estas bases de datos son de código abierto existen demasiados motores, pero la mayoría se encuentran en etapa de "proyecto", es decir, aún con modificaciones importantes e inestables, no han tenido mucha publicidad (carecen de confiabilidad y madurez) o no se ha visto una implementación importante de estas- que se consideran tienen características importantes y únicas para tener un abanico diverso y seleccionar apropiadamente una alternativa más adelante durante el proceso de elaboración del trabajo de grado.

Los motores se describen con la mayor cantidad de información que sea posible de acuerdo a los ítems de la Tabla 1.

| **Características** | | |
|---|---|---|
| Mantenibilidad | Características de la máquina | Instalación/Implementación |
| Usabilidad | Versión | Soporte y solución de inquietudes |
| Madurez | Documentación | Ventajas/Desventajas/Limitaciones |

**Tabla 2. Características NoSQL**

Características:

- Mantenibilidad: esfuerzo necesario para identificar y corregir errores, así como implementar nuevas funcionalidades.

- Usabilidad: interacción entre el sistema y el usuario, entendimiento y aprendizaje del usuario de cómo utilizar el motor de búsqueda. Si el sistema cuenta con una interfaz de usuario, disminuye el tiempo que toma un usuario inexperto en realizar actividades y es más atractivo para los clientes.

- Madurez: experiencia tanto del proveedor como del motor de búsqueda, ya que genera confianza y seguridad en los usuarios, indica el esfuerzo y desarrollo en las versiones.

- Características de la máquina: sistema operativo, versión y software necesario para el funcionamiento de la base de datos.

- Versión: todo programa o aplicación cuenta con varias versiones, en las que el proveedor modifica, corrige o implementa nuevas funcionalidades. Se encuentran versiones estables e inestables, se recomienda usar la versión estable ya que es más confiable, segura y en la cual se corrigieron la mayoría de los errores encontrados durante el proceso de pruebas.

- Documentación: Documentos de soporte como: manual de instalación y funcionamiento y librería de comandos que son permitidos por el motor de base de datos.

- Instalación/Implementación: cantidad de procedimientos y pasos que debe seguir el usuario para descargar los archivos y seguidamente instalarlos para hacer uso de la base de datos.

- Soporte y solución de inquietudes: creación de mesas de ayuda, foros y espacio para “Contáctenos” por parte del proveedor de la base de datos para publicar recomendaciones o dudas con respecto a alguna funcionalidad del motor.

- Ventajas/Desventajas/Limitaciones: características a resaltar de cada motor y posibles restricciones de uso que es importante tener en cuenta cuando se instale o implemente la solución.

La búsqueda se realiza en internet debido a que es un tema que pocos conocen y que es relativamente nuevo, por lo que encontrar información física es difícil; con el objeto de complementar la información obtenida se procederá a indagar en trabajos de grado, tesis o noticias sobre temas relacionados que puedan corroborar la veracidad de los datos encontrados en las diferentes páginas web durante la búsqueda.

Paralelamente a esta actividad se comienza a obtener contactos en diferentes empresas para la realización de una encuesta (correspondiente al objetivo específico 1) que será el núcleo del trabajo de grado sobre el que giren las decisiones a tomar de la alternativa NoSQL. La encuesta se realizó inicialmente en Microsoft Word 2007 y luego se adaptó a Google forms, un servicio de formularios que presta Google para la realización de encuestas por medio de internet.

La encuesta consta de 16 preguntas basadas en su mayoría en conocer qué bases de datos utiliza la empresa para la que trabaja el encuestado y las actividades que involucra poseerlas, como mantenimiento, políticas de uso, requisitos, entre otras; y la parte restante, en decisiones sobre el uso de los motores relacionadas también con las políticas de la empresa y conocimiento breve de NoSQL. Seguidamente se analizarán las respuestas dadas a cada pregunta en base a la muestra de empresas encuestadas y con

esta estructura se permitirá la realización de los objetivos siguientes e influencia sobre el concepto NoSQL.

Con las 2 formas de encuesta (archivo en Microsoft Word y Google forms) se enviaron las encuestas por medio de correos electrónicos a los contactos en 5 diferentes empresas.

Finalizando la obtención de los resultados de las encuestas, se inició la elaboración de la lista de chequeo correspondiente al objetivo específico 2 sobre el mismo archivo de Microsoft Word 2007 sobre el que se realiza el trabajo de grado. En esta lista se identifican los puntos clave a considerar de acuerdo a las respuestas obtenidas de los encuestados para optar por una alternativa de solución a los motores SQL.

Terminada la realización de las encuestas, se completa la lista de chequeo, sobre la cual las 4 bases de datos NoSQL se evalúan, de tal forma que se permita seleccionar y analizar por qué se escogió determinada opción, dando por completado el objetivo específico 3, y también el objetivo de este trabajo de grado.

# 3. ANÁLISIS DE LAS BASES DE DATOS ACTUALES Y REQUERIMIENTOS DE LAS EMPRESAS EN MEDELLÍN

**Objetivo específico 1**: Elaboración y realización de la encuesta relacionada con información general de los motores que utilizan las empresas y conocimiento sobre NoSQL. La encuesta consta de 16 preguntas de tipo excluyente, selección múltiple y abierta.

Se realizó un total de 5 encuestas dirigidas a personal encargado de la adquisición, administración y mantenimiento en diferentes empresas ubicadas en Medellín con alto grado de madurez, ya que son estas grandes empresas las que necesitan mejores tiempos de respuesta de las aplicaciones, mayor exigencia de calidad en los datos y decisiones más críticas en el uso de tecnologías. Adicionalmente la encuesta fue enfocada a este público debido a que las preguntas están relacionadas con los motores de cada empresa.

Una vez se obtienen los resultados de las encuestas, estos se documentan y se realiza el análisis respectivo para cada pregunta, de forma que se permita observar las necesidades generales de las empresas.

**Objetivo específico 2:** De acuerdo con los resultados obtenidos de la encuesta realizada, se realiza una lista de chequeo para evaluar las características y requerimientos de las empresas frente a las diferentes bases de datos NoSQL documentadas en este trabajo teniendo en cuenta los atributos de la tabla 2.

**Objetivo específico 3:** Cuando se tiene completa la lista de chequeo, se verifica cuál de las bases de datos de tipo NoSQL se puede considerar como mejor alternativa de acuerdo con los requisitos y necesidades de las empresas, para proceder a explicar y concluir sobre esta opción.

# 4. DISCUSIÓN DE RESULTADOS

## 4.1 RESULTADOS DE LA ENCUESTA

Como se mencionó anteriormente, se obtuvieron respuestas a las preguntas de la encuesta de 5 diferentes empresas localizadas en Medellín con alto grado de madurez y dentro de estas, dirigidas a personas con conocimiento en la toma de decisiones o administración sobre las bases de datos, proveyendo información útil para los demás objetivos específicos que se explicará más adelante.

Para identificar fácilmente los resultados obtenidos de las empresas por medio de las encuestas, se describirá la pregunta y se analizarán las respuestas como se muestra a continuación:

**Pregunta 1:** ¿Qué bases de datos ha adquirido/usado la organización para el almacenamiento de información en los últimos 5 años?

Se pretende como acercamiento inicial, conocer cuáles son los motores más utilizados. Según los resultados de las encuestas, las cuatro opciones propuestas eran las elecciones más frecuentes; además es normal para empresas grandes manejar no sólo una base de datos sino varias por lo que esta pregunta es de selección múltiple con múltiple respuesta. Como se puede ver en la figura 2, las bases de datos más utilizadas son Microsoft SQL Server y Oracle ya que todos los encuestados tenían por lo menos en su abanico de motores ambas opciones. Esto se puede reconocer debido a que son las nombradas en el medio para trabajos más robustos y debido a su integración y complementos con otras aplicaciones y servicios; aparte de estas dos opciones, un 60% de las empresas encuestadas también usan otras bases de datos como DB2, DB400 y ESSBASE, los dos primeros de IBM y el tercero de Oracle.

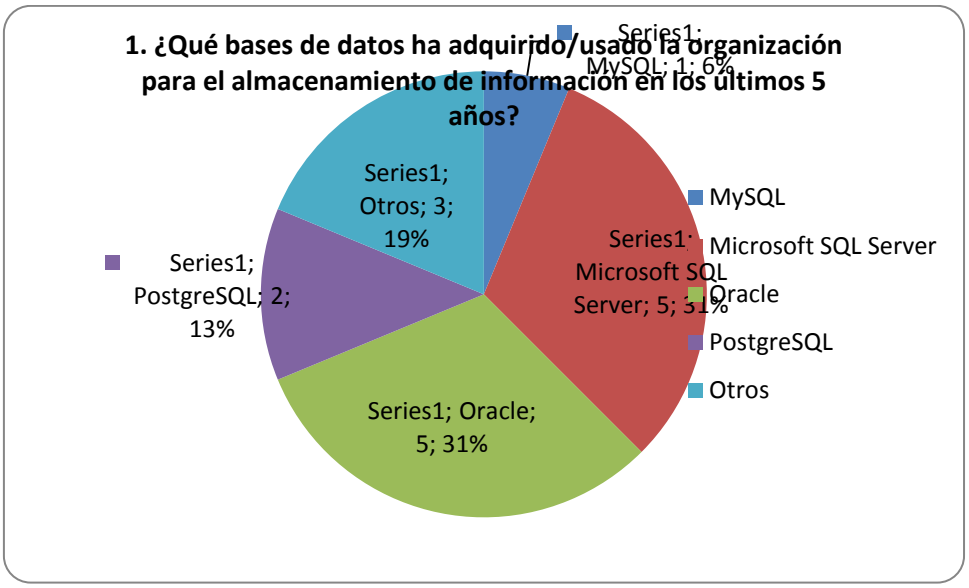


**Figura 2. Respuesta 1 de la encuesta**

**Pregunta 2:** ¿Qué características se tuvieron en cuenta para la utilización de las bases de datos anteriores?

Pregunta de selección múltiple con múltiple respuesta.

Generalmente las empresas no sólo tienen una especificación para comprar y adaptar un motor y según la figura 3, casi todas las características se tuvieron presentes, aunque la más sobresaliente fue la capacidad de integrarse con otras, seguida de la funcionalidad y velocidad en las transacciones. Se puede identificar que en el mundo de hoy muy pocas aplicaciones trabajan aisladas de los demás proyectos de la compañía, es por eso que el intercambio de datos, sincronización y verificación son importantes para mantener los información confidencial, integra y disponible. Como plus, las empresas también demandan funcionalidad y velocidad, es decir que sean flexibles, que permitan ejecutar varias actividades y que respondan ágilmente a las peticiones de las aplicaciones.

Dentro del proceso de respuestas a la presente pregunta, se generó una opción que no se tenía prevista y que cabe resaltarla ya que estará relacionada con el análisis que se realizará en una pregunta más adelante; la respuesta aportada está enmarcada en la adquisición de software y se refiere a la obligación de adquirir determinada base de datos en la compra de software. Estos se presentan cuando los proyectos de desarrollo no son personalizados de acuerdo a los requerimientos del cliente sino que se venden como productos sellados, de esta forma aquellos que se deben adaptar a las características son los compradores, en este caso las empresas.

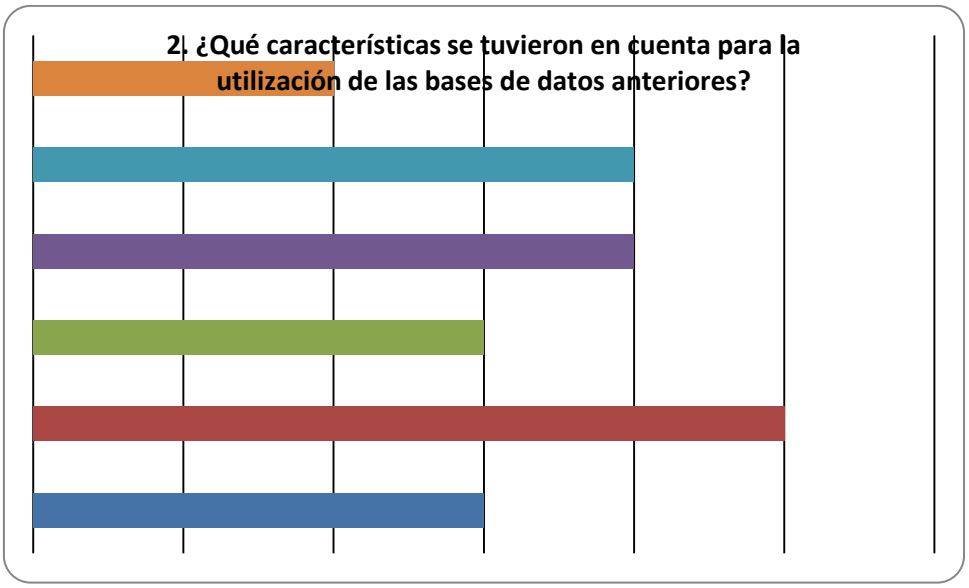


**Figura 3. Respuesta 2 de la encuesta**

**Pregunta 3:** ¿Qué factor es más importante cuando una aplicación hace uso de la base de datos?

Pregunta de tipo excluyente de dos posibles opciones para obtener cuál es el tema al que se le da más relevancia en un proyecto; esta está asociada tanto a nivel lógico como físico y aunque parece desviarse de las bases de datos y enfocarse más al desarrollo de un servicio o aplicación, es una decisión crucial dependiendo de la cantidad de recursos y tecnologías con los que cuente la empresa.

Generalmente la elección de tomar una u otra se realiza en las primeras etapas del ciclo del desarrollo de software según el modelo RUP para minimizar errores, costos y contratiempos en las etapas siguientes definiendo una arquitectura y rumbo a seguir durante el proyecto.

Los resultados para esta pregunta (Véase Figura 4) muestran que la mayoría de las empresas (80% de las encuestadas) tienen presente ante todo la velocidad en los tiempos de respuesta de las aplicaciones cuando se conectan a un motor de base de datos, es decir que “dejan” de lado los costos y la cantidad de almacenamiento enfocándose más en la optimización de consultas. Para lograr esto, se puede inferir que la estructura de cómo están los datos organizados las tablas no está normalizada, ya que de esta forma se asegura que los datos ya estén “preparados” y listos en vez de tener que realizar operaciones que ralentizan las salidas que espera el usuario.

Un claro ejemplo de modelos no normalizados es el de estrella, usado frecuentemente en soluciones de BI, que es propio para una empresa tenerlos ya que le ayudan a predecir comportamientos según datos históricos.

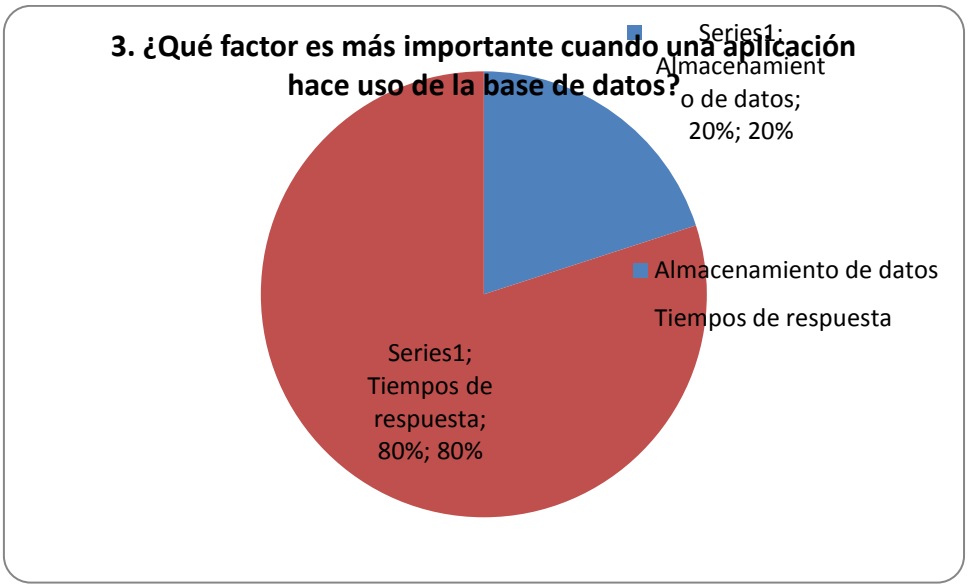


**Figura 4. Respuesta 3 de la encuesta**

**Pregunta 4:** ¿Qué funcionalidades son indispensables en un motor para que pueda hacer parte de la infraestructura de almacenamiento en la empresa?

Pregunta de selección múltiple con múltiple respuesta y seis posibles opciones. Las alternativas propuestas son usualmente las funcionalidades más visibles en los motores que utilizan las empresas y que se pueden visualizar en las respuestas dadas a la primera pregunta de la encuesta (Véase Figura 2).

La funcionalidad más seleccionada resultó ser el uso de vistas (*views*) como se ve en la figura 5. Las vistas funcionan como si se tomara una foto de los datos que existen en determinado momento pudiendo agrupar información de distintas tablas, y son útiles para controlar lo que puede ver un grupo de usuarios, reservando información confidencial que se hubiera podido encontrar si se daba acceso a una o varias tablas.

Una opción que es importante analizar es la respuesta libre por parte de los encuestados en la opción “Otros” en la cual otros atributos a tener en cuenta son: particionamiento, replicación, recuperación contra desastres, integración con otras aplicaciones y que tengan herramientas para la administración de los datos. Habiendo obtenido tan poca escogencia, son características transversales y necesarias para todas las empresas maduras; tomando como ejemplo la administración de los datos, se debe contar con que los datos son un activo más de la compañía y por lo tanto son valiosos y deberían ser confidenciales y exactos, además se debería realizar mantenimiento y auditorias para garantizar la calidad de la información, sin desaprovechar los beneficios que conlleva las otras cualidades anteriormente nombradas para facilitar el correcto funcionamiento, mantenibilidad y escalabilidad del motor a los administradores y auditores.


**Figura 5. Respuesta 4 de la encuesta**

**Pregunta 5:** ¿Sobre qué sistema operativo corre el servidor de bases de datos?

En esta pregunta es de respuesta libre ya que actualmente existen diversos sistemas operativos sobre los cuales se pueden montar las bases de datos. Estos fueron los resultados obtenidos:

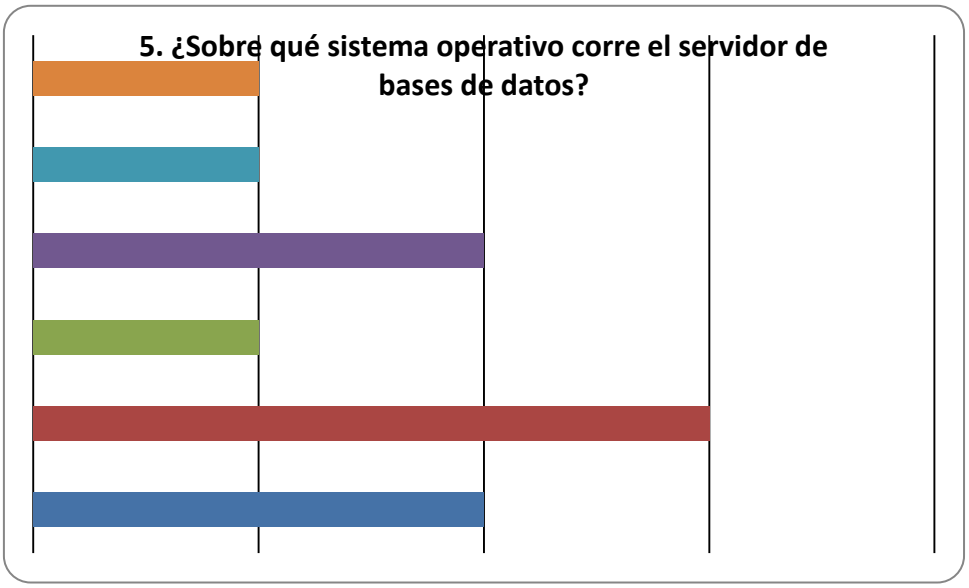


**Figura 6. Respuesta 5 de la encuesta**

De acuerdo a la muestra tomada, las variantes del sistema operativo Windows son las más usadas con un total de tres votos, y en un escalón más abajo se presentan las variantes del sistema operativos Unix y AIX –siendo AIX parte de Unix-; esto se puede

deber al desarrollo o compra de productos de software que sólo son compatibles con Windows, ya sea a nivel de formato de archivos o a integración con otras aplicaciones.

La información de la figura 6 se utilizará en los siguientes objetivos específicos de este trabajo de grado, de forma que se pueda identificar una base de datos NoSQL que sea compatible con la mayor cantidad de sistemas operativos anteriores.

**Pregunta 6:** ¿Cada cuánto tiempo realizan respaldos a la información contenida en la base de datos?

Pregunta de tipo excluyente con única respuesta de cuatro posibles y relacionada a la frecuencia con que se realizan procesos de back up. Las posibles de respuesta son:

- Varias veces a la semana
- Cada semana
- Cada mes
- Lapso superior a un mes

Se propusieron estos intervalos pero considerando la magnitud de las empresas se preveía que los resultados se ordenarían entre las dos primeras alternativas, aún así se dejaron las cuatro. Como resultado, todas las empresas encuestadas coincidieron en un solo punto: Realizar el procedimiento de respaldo varias a veces a la semana, y en algunas empresas respondieron además que lo realizaban diario.

Lo anterior demuestra la importancia y seguridad que se da al activo de la información en las organizaciones para prevenir desastres que entorpezcan o alteren los datos y que se deberá tener en cuenta en la selección del motor NoSQL.

**Pregunta 7:** ¿En alguna aplicación es necesaria la sincronización de datos en otro(s) servidor(es)?

Se redactó esta pregunta pensando en las características que tienen las empresas actualmente, normalmente se tienen varios medios de comunicación para el acceso a información o incluso dispositivos móviles desde el cual se escribe información a la base de datos, por lo que es indispensable sincronizar y verificar datos frente a otros servicios.

La modalidad de respuesta es de tipo excluyente (Si – No), en el que todas las empresas encuestadas respondieron “Si”. Se concluye entonces que al menos una aplicación necesita conectarse a un servidor y actualizarse los datos -similar a un proceso de replicación- en el que se identifican las versiones de los datos (origen y destino) y se determina en cuál versión deberían estar ambos.

**Pregunta 8:** ¿Tienen soporte por parte del proveedor de la base de datos?

Pregunta es de tipo excluyente (Si – No). Como respuesta nuevamente, todas las empresas dieron su opinión de sí tener soporte por el proveedor.

En todas las compañías maduras se tiene un área de informática (para desarrollo de software, evaluación, administración de aplicaciones, etc.) y cuando se presenta un problema con el motor o alguna operación en este, pueden contar con el personal interno, pero si el problema persiste, tienen soporte externo, que les da seguridad y solución. Es el caso en el que nuevas versiones de algún software aparecen y por ende se da por terminado el soporte a lanzamientos antiguos, y la empresa debe optar por 2 opciones: actualizarse a una más reciente para obtener respaldo del proveedor o continuar con la misma, pero en caso de fallo la empresa misma debe buscar por otros medios para dar por finalizado el problema.

**Pregunta 9:** ¿Cuánto es el costo mensual de mantenimiento y administración de una base de datos en un servidor local?

Pregunta de opción múltiple con única respuesta de tres posibles.

Observando la figura 7, las empresas encuestadas gastan más de $500.000 pesos colombianos en el mantenimiento de sus propios servidores de datos; de los cuales el 40% asigna hasta $1.000.000 de pesos y el 60% más de este.

El mantenimiento es realizado por administradores de las bases de datos que conocen su estructura interna de forma que sea posible mantenerla operando normalmente; las actividades que involucra son propias de cada motor como verificación de errores, actualización de índices sobre las tablas, cálculo de espacio libre, verificación de permisos, entre otras. Estas y todas las otras políticas internas de cada empresa para el mantenimiento necesitan la asignación de recursos (procedimientos, tecnologías, personas), y aunque según las respuestas de la pregunta dos de la encuesta (Véase Figura 3) el precio no es realmente importante frente a otras características, si facilita la realización y seguimiento de acciones preventivas y correctivas en las bases de datos.

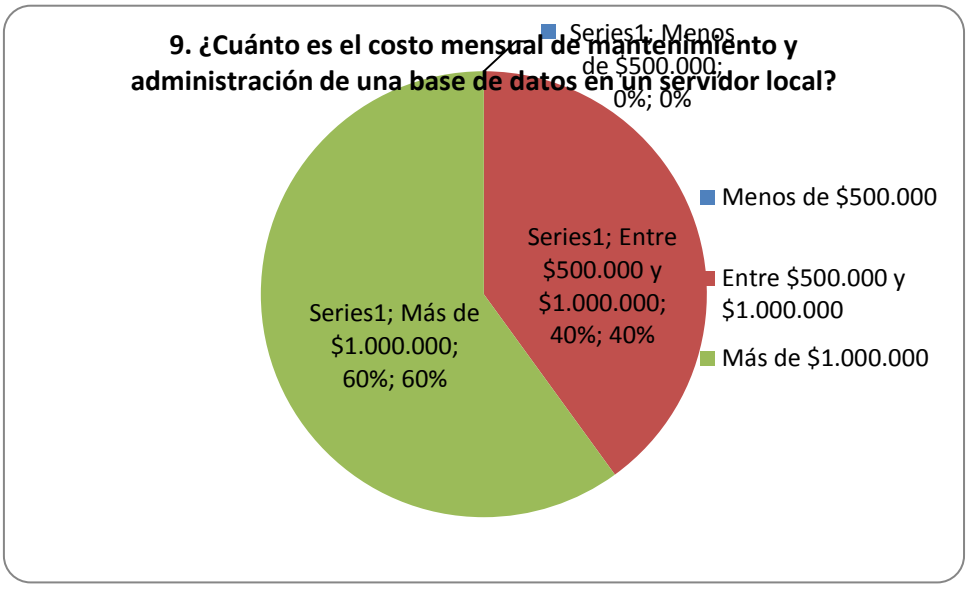


**Figura 7. Respuesta 7 de la encuesta**

**Pregunta 10:** Cuando sale al mercado una nueva versión del motor, ¿lo adquieren inmediatamente?

Pregunta es de tipo excluyente (Si – No).

El 80% de las empresas encuestadas respondieron que no lo adquirían mientras que un 20% si lo hacía como lo muestra la figura 8.

Esto no indica que contradice lo analizado en la pregunta 8 concerniente al valioso soporte que del proveedor para las empresas. Cuando se realiza un nuevo lanzamiento del motor, este permanece en una etapa beta para que los clientes la evalúen y exploren las correcciones realizadas y nuevas funcionalidades; además no indica que la versión del sistema de bases de datos vaya a quedar obsoleta o desatendida por el proveedor, en este caso si aún cumple con las características vistas anteriormente (Véase Figura 3) entonces se puede conservar por un poco más de tiempo.

Para el caso del 20% se podría afirmar que tiene recursos suficientes para estar a la vanguardia y prestar servicios más completos.

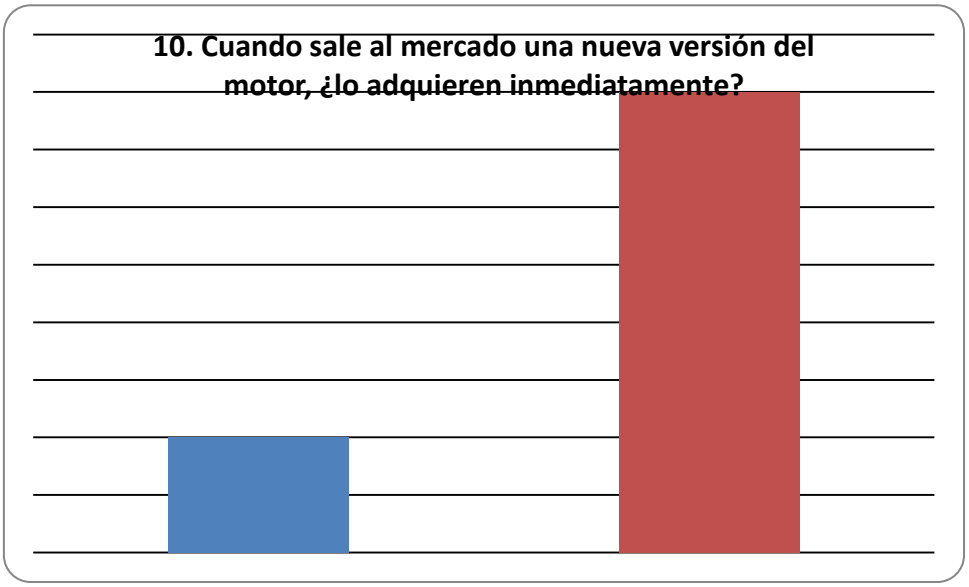


**Figura 8. Respuesta 10 de la encuesta**

**Pregunta 11:** ¿Las aplicaciones actuales son en su mayoría desarrollos internos o realizados por terceros?

Las posibles repuestas eran dos y sólo se podía seleccionar una:

- Desarrollo interno
- Desarrollado por terceros

La respuesta en común de todos los encuestados fue la segunda: tercerizar los desarrollos. Como se dijo antes, las empresas tienen sus áreas para el desarrollo y efectivamente realizan sus propios proyectos pero frecuentemente contratan uno o más terceros para diferentes actividades lo que le permite a la compañía enfocarse en otras soluciones.

En la contratación de terceros una ventaja es delegación del trabajo, dándoles instrucciones y condiciones aceptables para la entrega que deberán tener soporte sobre el proyecto realizado, pero implica una salida de dinero dependiendo de la magnitud y cubrimiento de actividades del ciclo de desarrollo de software; la otra alternativa es la compra de software que pueda existir en el mercado que no es personalizado y que trae consigo los requerimientos de instalación, como en el caso especial en la pregunta 2 de la encuesta.

**Pregunta 12:** De forma general, ¿qué consideraciones tienen presentes para la migración de los datos a una nueva versión u otro motor de base de datos?

Respuesta abierta y, como la pregunta lo dice, se pretende saber los requisitos para la migración. La migración en si es algo muy común para las áreas informáticas de las compañías y que se ha abordado anteriormente en las preguntas de la encuesta, por ejemplo cuando ya no se tendrá soporte por parte del proveedor, se procede a realizar la migración de los datos y normalmente se realiza sobre los productos del mismo proveedor.

Las respuestas obtenidas fueron las siguientes:

- Consistencia y compatibilidad: Asegurar que los datos quedarán íntegros y que la aplicación pueda usar normalmente a la base de datos.

- Estabilidad de la nueva versión: evaluar las características de la nueva versión para identificar posibles fallas tanto internas como externas a los motores de base de datos.

- Continuidad de la funcionalidad: evaluar los riesgos que pueden ocurrir cuando se migra los datos para asegurar que las aplicaciones puedan seguir funcionando correctamente.

- Soporte: tiene que ver con el soporte por parte del proveedor tanto del motor actual, ya que se ha dicho que si aún se tiene y no existe necesidad, las empresas prefieren quedarse en este estado. También tiene que ver con el soporte y condiciones que se tendrán una vez se migre a otro sistema.

- Esfuerzo: cantidad de recursos y tiempo necesarios para implementar la migración

**Pregunta 13:** ¿Existe alguna política en la empresa que obligue a almacenar la información en instalaciones de la empresa?

Pregunta de tipo excluyente (Si – No) para identificar las directrices de almacenamiento dentro de la misma compañía.

Según la figura 9, el 60% de las empresas obligan a que los datos permanezcan en servidores propios, siendo estos administrados por la empresa misma. El restante 40% puede que tengan una política más liberal y que tengan información mantenida por un tercero. En algunos casos los nuevos modelos de bases de datos proveen el servicio completo de control y mantenimiento, que se tendrá en cuenta más adelante para la selección de bases de datos que cumplan con los requisitos de las compañías.

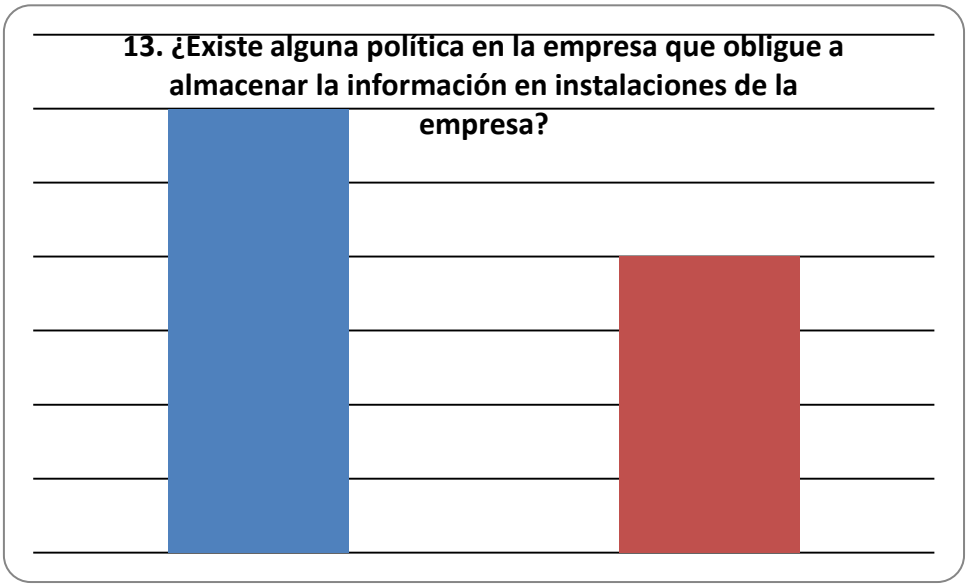


**Figura 9. Respuesta 13 de la encuesta**

**Pregunta 14:** ¿Estaría dispuesto a utilizar algún sistema de almacenamiento ubicado en instalaciones de un proveedor especializado que brinde garantías de seguridad disponibilidad e integridad?

Pregunta es de tipo excluyente (Si – No) y está relacionada con la pregunta anterior, para conocer la disposición sobre utilizar servicios de terceros para el almacenamiento de los datos.

Con la respuesta anterior se identificó que 60% tenía directrices explícitas sobre el almacenamiento local, pero para esta pregunta un 80% aceptaría usar servicios externos siempre y cuando aseguren los atributos de seguridad (confidencialidad, integridad y disponibilidad). Un 20% sin embargo no estarían dispuestos y preferirían seguir administrando los datos internamente en instalaciones de la compañía.

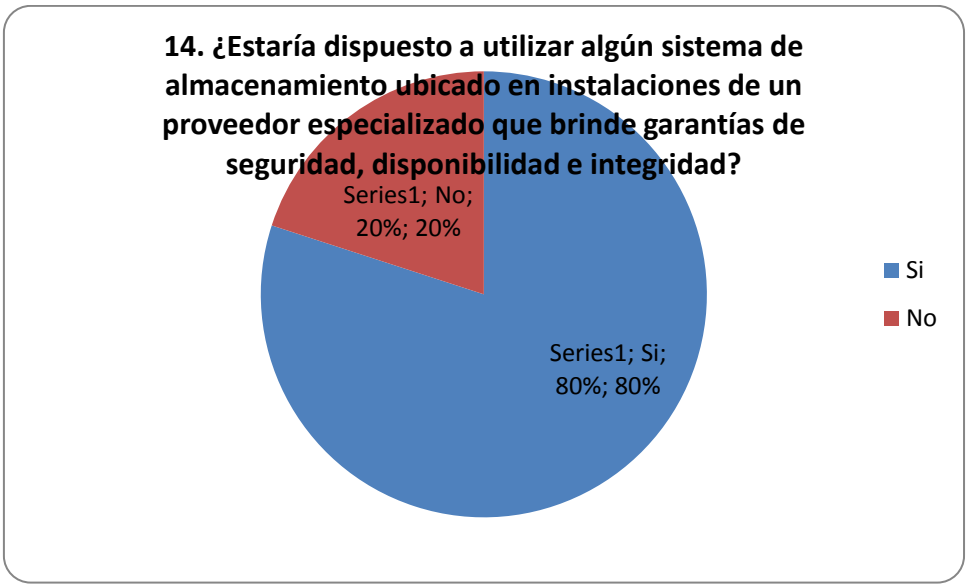


**Figura 10. Respuesta 14 de la encuesta**

**Pregunta 15:** ¿Conoce el término NoSQL?

Esta pregunta de tipo excluyente (Si – No) ya no está relacionada con las bases de datos o su administración, sino con el conocimiento por parte de los encuestados respecto al concepto NoSQL.

De la figura 11, se puede decir que un 60% de los encuestados ha escuchado el término mientras que un 40% no lo conoce. Esto puede indicar el grado de tecnologías que presentan las empresas con los temas actuales y que se han comenzado a difundir en los medios bien sea por medio de internet o por conferencias.

La ventaja que presenta conocer este u otro concepto actual para cualquier ámbito de conocimiento es que facilita el entendimiento y razonamiento para optar por una tecnología cuando se den varias opciones de cambio.

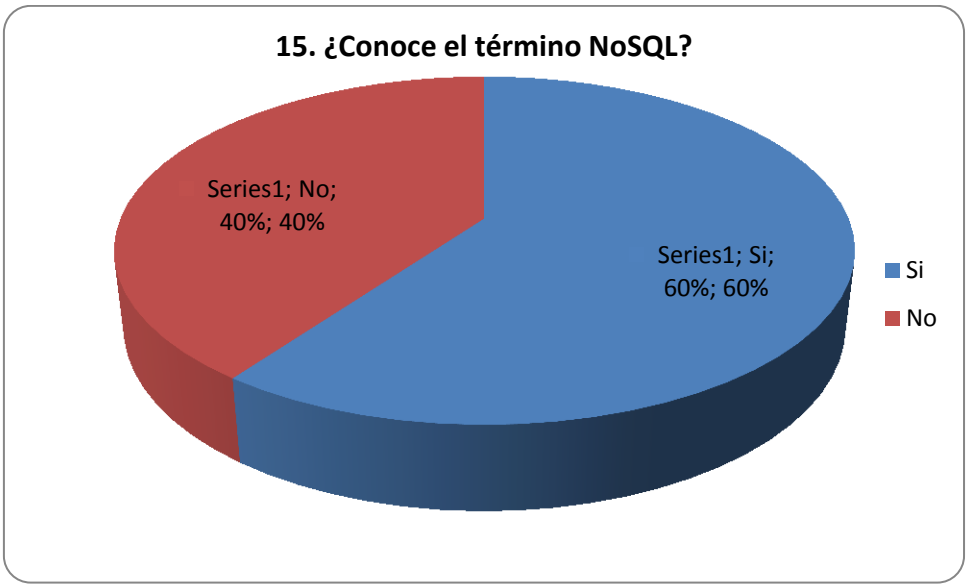


**Figura 11. Respuesta 15 de la encuesta**

**Pregunta 16:** ¿Conoce alguna de las siguientes bases de datos?

Pregunta de selección múltiple con múltiple respuesta sobre el conocimiento de bases de datos NoSQL.

Se escogieron estos motores ya que son, entre los tantos que existen, los más reconocidos por decirlo de alguna manera: SimpleDB y Dynamo de Amazon; BigTable de Google y Cassandra de Facebook.

Según la figura 12, las dos empresas que, en la respuesta anterior las empresas que no conocían el término respondieron que no conocían ninguna base de datos, es decir que entre las 3 restantes que sí lo conocían completaron las selecciones de las cuatro opciones; siendo Dynamo la que más reconocían.

Como el análisis anterior, representa simplemente el conocimiento frente a este tema y se analizará más adelante las razones para seleccionar un motor NoSQL.

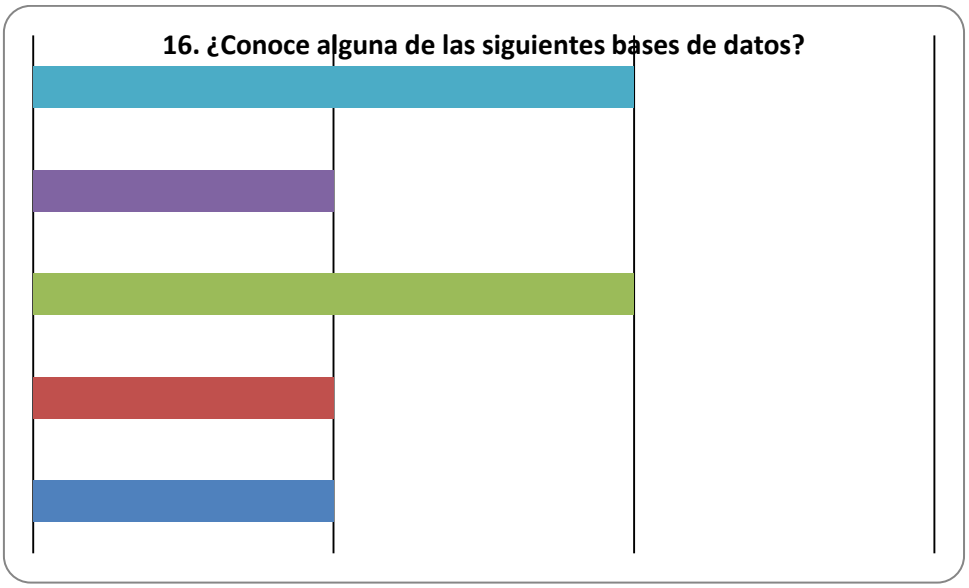


**Figura 12. Respuesta 16 de la encuesta**

## 4.2 LISTA DE CHEQUEO

Esta actividad se realizó con la información obtenida de las respuestas y análisis de la encuesta anterior (Numeral 4.1).

A continuación se mostrará una lista de chequeo con la estructura Características vs. Bases de datos NoSQL: las características son las necesidades de las empresas según el párrafo anterior, y los motores NoSQL son los que se nombraron y describieron en numeral 1.3.1.

| | | **Bases de datos NoSQL** | | | |
|---|---|---|---|---|---|
| | Características | Cassandra | MongoDB | Neo4j | DynamoDB |
| **1** | ¿Tiene marco de trabajo gráfico? | Si | Si | Si | Si |
| **2** | ¿Se tiene soporte por parte del proveedor? | No | No | Si | Si |
| **3** | ¿Se puede integrar con otras aplicaciones? | Si | Si | Si | Si |
| **4** | ¿Optimiza el tiempo de repuesta en las transacciones? | Si | Si | Si | Si |
| **5** | ¿Tiene una versión estable de por lo | Si | Si | Si | Si |

| | | | | | |
|---|---|---|---|---|---|
| | menos este año? | | | | |
| **6** | ¿Almacena los datos de forma óptima? | Si | Si | Si | Si |
| **7** | ¿Contiene procedimientos almacenados, vistas, triggers, etc.? | Si | No | Si | Si |
| **8** | ¿Se pueden particionar las tablas que contienen los datos? | Si | Si | Si | Si |
| **9** | ¿Permite la replicación de los datos? | Si | Si | Si | Si |
| **10** | ¿Puede funcionar bajo sistema operativo Windows? | Si | Si | Si | Si |
| **11** | ¿Puede funcionar bajo sistema operativo AIX? | No | No | Si | No |
| **12** | ¿Se pueden realizar procesos de back up varias veces a la semana? | Si | Si | Si | Si |
| **13** | ¿Es posible sincronizar los datos con otro servidor? | Si | Si | Si | Si |
| **14** | ¿Se asegura la integridad, confidencialidad y disponibilidad de los datos? | Si | Si | Si | Si |
| **15** | ¿Tiene medidas de aseguramiento de recuperación de desastres? | Si | Si | Si | Si |
| **16** | ¿Tiene herramientas que apoyen la administración de los datos? | Si | Si | No | Si |
| **17** | ¿Los datos son almacenados en servidores propios del usuario? | Si | Si | Si | No |
| **18** | ¿Es usuario es el encargado de la gestión y mantenimiento de los datos? | Si | Si | Si | No |

**Tabla 3. Lista de chequeo**

A continuación se dará una explicación para cada una de las preguntas anteriores de forma que se pueda entender por qué se consideró y que bases de datos cumplen ese requisito:

**Pregunta 1:** ¿Tiene marco de trabajo?

De acuerdo al análisis realizado de las respuestas de la segunda pregunta de la encuesta, la usabilidad –que incluye las GUIs- no fue una consideración tan importante para las empresas como lo fue la integración, pero si ayuda a agilizar el proceso de creación y modificaciones en la base de datos y hace más amigable la interacción del usuario con el sistema.

Todas las bases de datos NoSQL de alguna forma cumplen este objetivo, ya sea porque tienen herramientas realizadas por la misma empresa que desarrolló el motor o implementadas por terceros.

**Pregunta 2:** ¿Se tiene soporte por parte del proveedor?

Requisito que es necesario por todas las empresas para tener asesoría y respaldo de la funcionalidad operativa de la base de datos.

Para Cassandra y MongoDB no existe soporte alguno, lo único que se ofrecen son las ayudas por medio de chat, foros, manuales, etc.

Para Neo4j, la versión que no tiene soporte es la *open source* (versión comunitaria), las otras dos alternativas implican una compra y acuerdo de servicio, lo que conlleva a un soporte.

DynamoDB cuenta con AWS Support, un canal de soporte personalizado cualquier día de la semana y tendrá un costo dependiendo del tipo de servicio que se adquiera (Basic, Developer, Business, Enterprise) cada uno incluyendo más medios de comunicación y respuestas más rápidas

**Pregunta 3:** ¿Se puede integrar con otras aplicaciones?

A simple vista, de los sistemas NoSQL escogidos, la opción que parece permitir integración con otras aplicaciones o servicios es DynamoDB, ya que pertenece al grupo de servicios AWS y permite interactuar con otras bases de datos y herramientas de desarrollo como Eclipse, pero Neo4j tiene algunas integraciones con otras como Gelphi, una plataforma de visualización y exploración interactiva para todas las clases de redes y sistemas completos, dinámicos y gráficos jerárquicos. (Gelphi Consortium)

Cassandra y MongoDB también cuentan con integraciones un poco más técnicas con otros sistemas y se pueden considerar como mejoras o modificaciones entre el motor NoSQL y el otro software que hacen uso de los APIs para almacenar información e indexar datos y convertir objetos de JavaScript por ejemplo.

**Pregunta 4:** ¿Optimiza el tiempo de repuesta en las transacciones?

Un 80% de las empresas encuestadas respondieron que era preferible respuestas rápidas a preocuparse por el espacio de almacenamiento usado. En este contexto, las bases de datos NoSQL se diseñaron exactamente con este propósito, que fueran escalables y mantenibles, por esto muchas están basadas en algoritmos de funcionalidades que se ha visto que han tenido éxito como las implementaciones internas de Google y Amazon. De

acuerdo con lo anterior, todas opciones son consideradas optimizadas, claro que algunas están dirigidas conceptos y públicos diferentes.

**Pregunta 5:** ¿Tiene una versión estable de por lo menos este año?

En lo que lleva de este año, se debe esperar que se haya lanzada al público una versión estable, de lo contrario indica que el proyecto está atrasado o parado y se puede considerar un riesgo implementar el motor debido a errores futuros que no serán corregidos o tomarán una larga espera, además muestra el interés y esfuerzo por darse a conocer en el mercado.

En este caso, los 4 motores seleccionados cumplen este requisito:

- Cassandra: versión estable1.1.5
- MongoDB: versión estable 2.0.6
- Neo4j: versión estable 1.8
- DynamoDB: se hizo público a comienzos de este año, y aún se encuentra en la versión beta.

**Pregunta 6:** ¿Almacena los datos de forma óptima?

Como la pregunta 4, NoSQL cambia los paradigmas de las bases de datos SQL y permite mayor flexibilidad y agilidad. Cada motor NoSQL tiene su propia distribución de los datos dependiendo del público objetivo al que apuntan como el caso de los tipos mencionados: alternativas basadas en *key-value* y orientadas a grafos logrando mejores resultados en el modelo de almacenamiento.

**Pregunta 7:** ¿Contiene procedimientos almacenados, vistas, *triggers*, etc.?

Asociada a la pregunta 4 de la encuesta realizada a las empresas, las funcionalidades que se evalúan son: procedimientos almacenados, vistas, *triggers*, funciones y eventos programados.

MapReduce permite utilizar procesos *batch* y operaciones agregadas

Cassandra: análogo a los *trigger* SQL, contiene procedimientos llamados Asynchronous triggers para responder a ciertos eventos en la base de datos. Para los demás elementos Cassandra se ha integrado con funcionalidades como Apache Hadoop y Apache Hive –entre otras más- para usar funciones y facilitar la extracción, transformación y carga de datos (proceso ETL), además las sentencias son ejecutadas por medio de MapReduce para la realización de tareas.

MongoDB: no tiene las funcionalidades comúnmente encontradas en los sistemas SQL, pero tiene diferentes formas para ejecutar actividades y funciones como MapReduce por medio de JavaScript en el servidor. El uso de *triggers*, no está implementado aún, pero parece que se buscará un acercamiento.

Neo4j: como sus anteriores contiene componentes que realizan tareas similares: para el caso de *triggers*, cuenta con funciones para registrar las tareas por medio de controladores de eventos para cambios en los nodos; de igual forma estas funciones son similares a las funciones SQL.

En los sistemas Cassandra y Neo4j, la posibilidad de programar las propias funciones con características de *trigger*, permiten agendar cuándo ejecutar estas por lo que cumplirían con la creación de eventos programados.

DynamoDB: usando Amazon Elastic MapReduce (EMR) facilita el desarrollo de scripts en diferentes lenguajes de programación para la creación de funcionalidades y consultas a la base de datos.

Con las descripciones anteriores se tomará que Cassandra, Neo4j y DynamoDB soportan las características o funcionalidades mencionadas en la pregunta.

**Pregunta 8:** ¿Se pueden particionar las tablas que contienen los datos?

La partición de las tablas o los conceptos análogos en los sistemas NoSQL se permiten a través del *sharding* o fragmentación, esto permite reducir el número de filas por tabla, lo que incrementa el rendimiento de las transacciones cuando se realizan operaciones sobre ellas. Esto hace que el concepto NoSQL sea escalable horizontalmente adquiriendo más nodos en vez de mayores recursos para un solo servidor o nodo.

**Pregunta 9:** ¿Permite la replicación de los datos?

Cassandra almacena copias, llamadas replicas, de cada fila basada en la clave de la fila, este proceso se realiza cuando se está creando el *keyspace* –elemento que contiene las "tablas" de la base de datos-

MongoDB soporta la replicación asíncrona entre varios servidores en caso de fallas y para redundancia de datos, de modo que es posible proveer un servicio continuo. De este modo un servidor escribe solamente mientras los otros están atentos para la lectura y se asegura que los datos estén actualizados como se observa brevemente en la figura 13.

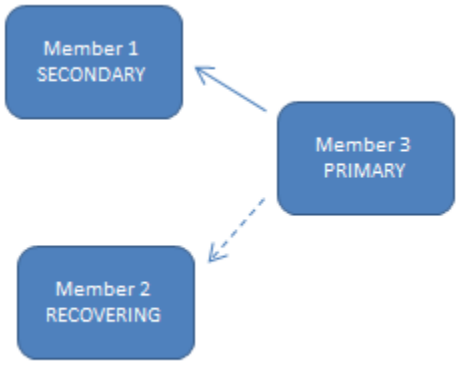


**Figura 13. Replicación MongoDB**

Imagen tomada de: http://www.mongodb.org/display/DOCS/Replication

Neo4j funciona como el caso anterior, con un concepto de tipo maestro esclavo en los *clusters* en el que uno escribe y el otro escucha coordinando y replicando los datos.

DynamoDB replica los datos de manera automática y sincronizada entre tres zonas de disponibilidad de una región a fin de ofrecer un alto nivel de disponibilidad y durabilidad de los datos frente a fallos de la máquina. (Amazon.com Inc.)

**Pregunta 10:** ¿Puede funcionar bajo sistema operativo Windows?

Según las repuestas dadas a la pregunta cinco de la encuesta, el sistema Windows fue la opción más elegida y la que tiene compatibilidad las base de datos NoSQL, de igual forma, no en todas estas se especifica exactamente cuál versión de Windows es la que soportan, por ejemplo Neo4j funciona mínimo bajo Windows XP.

Con esto se debería probar primero si efectivamente funciona para la versión del sistema que se tenga de Windows para verificar el correcto funcionamiento y evitar futuras complicaciones.

**Pregunta 11:** ¿Puede funcionar bajo sistema operativo AIX?

Se consideró poner este requisito, ya que fue el segundo sistema operativo más usado por las empresas. Unix también quedó en el segundo lugar, pero Unix en sí tiene varios sistemas como FreeBSD, Solaris e incluso AIX es un sistema Unix, por lo que no se consideraron todas estas variantes.

Con el motor Cassandra, no se encontró información que demostrara que se podía instalar, por lo que no cumpliría; MongoDB, no lo soporta; Neo4j no lo indica explícitamente, pero parece que un archivo de configuración tiene una condición para evaluar donde se encuentran los archivos correspondientes a JRE en el sistema de IBM; por último la administración de DynamoDB funciona como un servicio suministrado por Amazon, no se tiene que instalar o implementar nada directamente en la máquina, sólo acceso a internet a través de un navegador web para gestionar los diferentes servicios de AWS por lo que no podría acceder desde este sistema operativo. En nuestra opinión si es solamente para la gestión y comportamiento del motor, se puede ingresar a través de otros navegadores que sí son permitidos.

**Pregunta 12:** ¿Se pueden realizar procesos de back up varias veces a la semana?

Los procesos de respaldo son implementados o controlados cuando se comienza a crear la base de datos y son constantemente usados en la replicación. En las versiones empresariales de Neo4j y de la aplicación OpsCenter para Cassandra es posible realizarlos en vivo.

**Pregunta 13:** ¿Es posible sincronizar los datos con otro servidor?

Para todos los motores seleccionados es posible este atributo, ya que hace parte de la funcionalidad de la replicación.

**Pregunta 14:** ¿Se asegura la integridad, confidencialidad y disponibilidad de los datos?

Los tres conceptos para la seguridad se cumplen en las bases de datos, aunque se debe tener en cuenta que en algunos casos como la versión comercial Enterprise de Neo4j se ofrece mayor disponibilidad que sus otras dos.

Confidencialidad: autorizaciones y autenticaciones como Cassandra; con algoritmos para mantener los datos seguros como lo hace DynamoDB; autenticación y modos seguros en las conexiones como MongoDB; y aseguramiento de puertos y conexiones remotas, soporte HTTPS y control de accesos de rangos de IP y URLs como Neo4j.

Integridad y disponibilidad: a través de la fragmentación (denominado *sharding*) y la replicación de datos.

**Pregunta 15:** ¿Tiene medidas de aseguramiento de recuperación de desastres?

Esta es una funcionalidad que les permite a las empresas seguir operando y ofreciendo sus servicios y a su vez evitando la pérdida de los datos como se ha mencionado anteriormente.

En este caso la función de replicación lo cubre, ya que permite, en caso de error, seleccionar otro nodo que tiene los mismos datos para que continúe con las actividades normales como se vio en la figura 13.

**Pregunta 16:** ¿Tiene herramientas que apoyen la administración de los datos?

Esta característica les permite a las empresas analizar los datos y el comportamiento de la base de datos para tomar mejores decisiones en el futuro.

Cassandra: cuenta con aplicaciones de terceros para la administración que permiten la gestión y análisis de los datos.

MongoDB: al igual que Cassandra, cuenta con herramientas de terceros para el monitoreo de la información. La mayoría es de tipo *open source*.

Neo4j: No se encontraron aplicaciones que apoyen la gestión de los datos.

DynamoDB: se logra a través de la misma herramienta que proporciona Amazon denominada AWS Management Console para visualizar los datos, monitorear los recursos de escritura, lectura, red y almacenamiento, y controlar el estado de las bases de datos.

**Pregunta 17:** ¿Los datos son almacenados en servidores propios del usuario?

Para los motores Cassandra, MongoDB y Neo4j, el usuario se debe encargar de descargar los archivos fuentes para la instalación e ingreso de datos, por lo que son contenidos en equipos de cliente.

Contrario ocurre con DynamoDB, que son guardados en los servidores de Amazon.

**Pregunta 18:** ¿Es usuario es el encargado de la gestión y mantenimiento de los datos?

Esta pregunta complementa la anterior, ya que generalmente si un proveedor ofrece los servicios de almacenamiento, provee también servicios adicionales como ocurre con DynamoDB. Amazon controla la cantidad de los recursos según los solicite la aplicación, así como mantener operando continuamente la base de datos, quitándole esta carga al cliente. Para los otros tres motores, el usuario se debe encargar de tener las características necesarias de software y hardware para la replicación y configuración de los servidores.

## 4.3 ANÁLISIS ALTERNATIVA NOSQL

Esta sección se da como resultado del análisis y resultados realizados en el numeral anterior de acuerdo con la lista de chequeo de las características que deberían tener las bases de datos NoSQL.

Se describirá cuáles son los motores que cumplen con la mayor cantidad de requisitos o características que buscan las empresas en los sistemas de bases de datos y las consideraciones o limitaciones que pueden presentar.

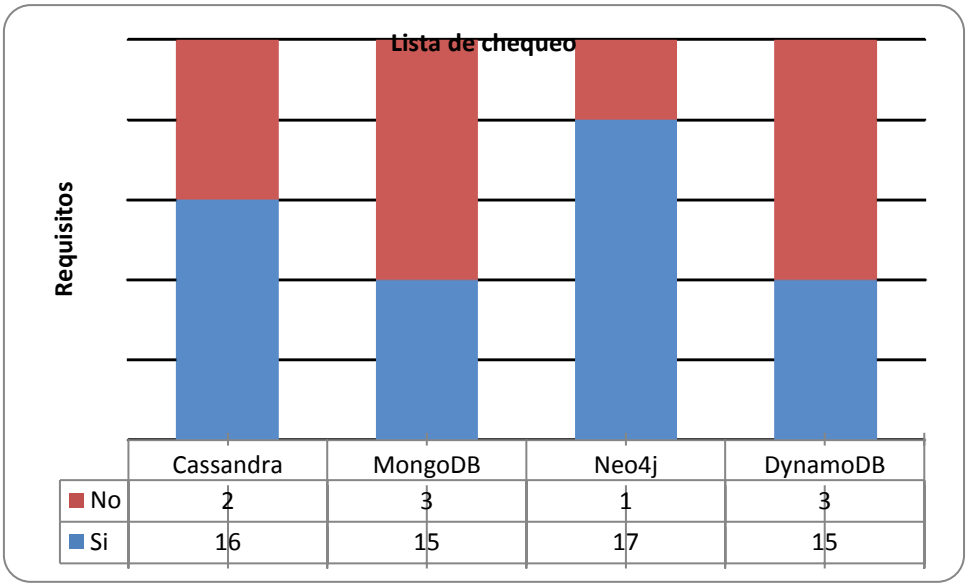


**Figura 14. Resultados lista chequeo**

Para la alternativa NoSQL, se considerará de acuerdo con la figura anterior, la que tenga más respuestas “Si”, ya que satisfacen los requisitos analizados después de realizar las encuestas a las organizaciones.

Según esto el orden sería el siguiente:

1. Neo4j

2. Cassandra
3. DynamoDB
4. MongoDB

DynamoDB, de Amazon, aunque obtuvo un tercer lugar, consideramos que debería ser el primer acercamiento NoSQL para las empresas ya que es un concepto que aún no es muy reconocido y con esta se facilita la operación y mantenimiento de los elementos del motor, contienen un costo directo asociado para utilizarlo, pero no sería un problema debido a las ventajas de integración, soporte, respaldo, entre otros atributos vistos en los numerales anteriores.

Se debe tener en cuenta que en este sistema los datos son almacenados y administrados por un tercero, es decir que para las políticas de la compañía en la cual se restringe el almacenamiento externo, no sería posible; pero sí sería viable y posiblemente a considerar como opción en aquellas que respondieron afirmativamente a la pregunta número 14 de la encuesta (80% de los encuestados), en la que se preguntaba sobre la disposición a utilizar instalaciones de un proveedor con garantías de confidencialidad, integridad y disponibilidad sobre los datos.

Como segunda alternativa a considerar para las empresas que no sea posible almacenar ninguna información fuera de la empresa es Neo4j, preferiblemente la versión Empresarial que tiene soporte y mayor disponibilidad, aunque es de resaltar que esta solución NoSQL está más enfocada a actividades del negocio donde se utilicen nodos e interacciones entre estos como lo son los grafos, de igual forma el resultado o el sistema con el que interactúa el usuario final son gráficas de datos geoespaciales. Para las compañías cuyo propósito no es el anteriormente mencionado, esta base de datos no es una opción viable.

Según la figura 14, la opción que cumple con la mayor cantidad de requisitos después de Neo4j es Cassandra, con propósitos más generales. Este motor es administrado y controlado completamente por el usuario y todas las responsabilidades de configuración, integración y mantenimiento deben ser realizadas por el cliente. Se debe considerar como riesgo el hecho de que no tenga soporte asociado.

Finalmente el orden que consideramos es el siguiente:

1. DynamoDB como primer acercamiento y entendimiento NoSQL.
2. Neo4j para gráficos con nodos.
3. Cassandra o MongoDB, optando por Cassandra que cumplió más requisitos.

De forma resumida y gráfica se puede identificar lo siguiente:

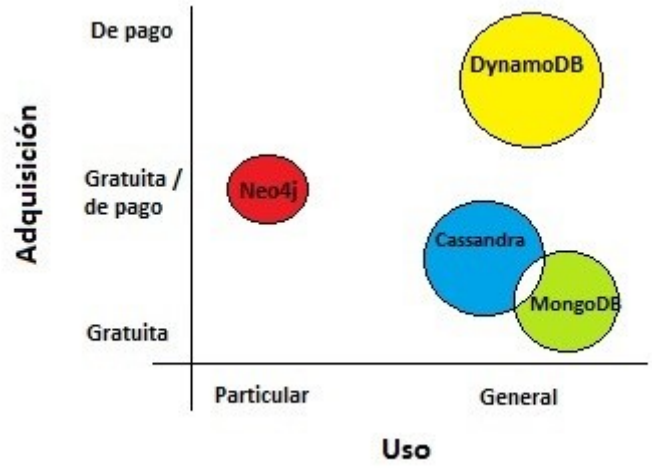


**Figura 15. Características bases de datos**

Esta figura representa gráficamente el análisis desarrollado anteriormente sobre cuál base de datos NoSQL deberían utilizar las empresas: DynamoDB, Cassandra y MongoDB están más concentradas en ofrecer funcionalidades generales, además se diferencian las modalidades de adquisición de los cuatro motores considerados para este trabajo de grado, desde gratuita hasta pagada, considerando también las dos licencias de Neo4j; se debe tener en cuenta que la adquisición -como se revisó en capítulos anteriores- está asociada al valor agregado y características adicionales que los distribuidores ofrecen de sus bases de datos.

Depende entonces de las políticas internas de la empresa, así como de la orientación a la solución que se necesita, para hacer un buen uso de los sistemas NoSQL. Igualmente se debe tener entendimiento en varios lenguajes de programación, ya que la mayoría son compatibles con estos motores y una buena gestión de recursos para embarcarse en los nuevos modelos de las bases de datos y sacarle provecho a las ventajas que conllevan.

# 5. CONCLUSIONES Y CONSIDERACIONES FINALES

Las conclusiones y consideraciones descritas a continuación son el fruto del análisis realizado a las encuestas sobre información general de las bases de datos usadas por las empresas y conocimiento sobre los motores NoSQL, así como la investigación de los cuatro sistemas evaluados como alternativas.

Las empresas generalmente prefieren utilizar bases de datos que se integren con otros servicios para facilitar el intercambio de datos y aprovechamiento de herramientas para la minería de datos, servicios web y procesos de ETL.

Ante todo, el tiempo de respuesta para el uso entre la aplicación y el motor de base de datos debe ser mínimo sin "importar" el tamaño del almacenamiento de los datos.

Entre las funcionalidades que debe tener un motor de base de datos en las empresas en cuanto a procedimientos, funciones, vistas, *triggers* y eventos programados, casi ninguna de tipo NoSQL lo cumple como frecuentemente se utiliza en los sistemas SQL, sino que presentan acercamiento y versiones análogas, por lo que habría que tener en cuenta los cambios e implementaciones de cada motor.

Entre los sistemas operativos más comunes para instalar los sistemas NoSQL se encuentran Windows y Linux. Aquellos que lo deseen instalar en otros tipos deben considerar la compatibilidad entre sistema operativo – base de datos.

Casi todas las bases de datos NoSQL están orientadas inicialmente a grandes segmentos del mercado de los sistemas operativos: generalmente están disponibles para sistemas Windows, Linux y Mac OS y poca o nula compatibilidad con otros Unix.

Los datos son un activo importante para la compañía que deben ser protegidos, por ende se deben realizar procesos de back up varias veces a la semana.

La sincronización entre aplicaciones es normal hoy en día, desde servicios en un mismo servidor hasta conexiones entre dispositivos móviles y los servidores de datos; se debe entonces realizar una gestión segura y correcta de los datos.

El soporte es indispensable para los motores de base de datos en las compañías y les permite mitigar el riesgo y problemas concernientes al motor.

Al ser la mayoría de las soluciones NoSQL de tipo *open source*, no presentan soporte por parte del proveedor y se debe considerar como un posible riesgo de adoptarlas.

Las nuevas versiones de las bases de datos tienen un proceso de acogida lento, ya que aún no existe la necesidad de adquirirlas o se espera primero que se identifiquen y corrijan errores de estas antes de implantarse; aunque eventualmente se deberían implementar por la vigencia del soporte.

El outsourcing les permite a las organizaciones enfocarse en los procesos *core* del negocio y es una modalidad muy utilizada, incluso más frecuente que desarrollar internamente.

Migrar es un proceso que toma tiempo e incluye desde la planeación para la identificación de riesgos, esfuerzo y aprovechamiento de nuevas funcionalidades hasta el traspaso de los datos como tal.

Según las directrices de la empresa, algunas están obligadas a almacenar los datos internamente pero estarían dispuestas a que el tercero los guarde siempre y cuando cumplan ciertos requisitos de confidencialidad, integridad y disponibilidad.

NoSQL es un concepto que aún en el medio empresarial no es muy reconocido, lo que dificultaría la introducción de este tipo de bases de datos.

La mayoría de las empresas tienen directrices que obligan a mantener los datos almacenados sus instalaciones, pero estarían dispuestas a que un tercero los almacene siempre que protejan y aseguren la información.

NoSQL cambia los paradigmas de los sistemas SQL y permite un almacenamiento y tiempos de respuesta más rápidos y optimizados de forma que sea escalable.

Algunas bases de datos NoSQL presentan varias versiones al público, siendo las de pago las que mayores beneficios traerían para las empresas, ya que cuentan con acuerdo de servicios y soporte.

Las alternativas NoSQL seleccionadas a evaluar durante este trabajo de grado cumplieron la mayoría de los requisitos que las empresas hoy en día buscan.

Los motores que son *open source* buscan atraer usuarios que aporten ideas y funcionalidades nuevas.

Las bases de datos NoSQL están en constante implementación y actualización y se pueden encontrar en aplicaciones o servicios que usamos a diario sin darnos cuenta.

Introducir un sistema NoSQL, al igual que cualquier otro componente a la arquitectura de una empresa, exige una ardua gestión del cambio y un proceso constante de adquisición de conocimiento por parte de los empleados y consideración de los riesgos.

Se debe considerar el objeto de la empresa o las características de la aplicación o servicio para realizar una acertada elección de las bases de datos.

# BIBLIOGRAFÍA


10gen. (18 de Octubre de 2011). *What is MongoDB?* Recuperado el 2012 de Agosto de 3, de 10gen: http://www.10gen.com/what-is-mongodb

Amazon.com Inc. (s.f.). *¿Qué es AWS?* Recuperado el 9 de 10 de 2012, de Amazon Web Services: http://aws.amazon.com/es/

Amazon.com Inc. (s.f.). *Amazon DynamoDB.* Recuperado el 21 de 10 de 2012, de Amazon Web Services: http://aws.amazon.com/es/dynamodb/

Amazon.com Inc. (s.f.). *Amazon Elastic MapReduce*. Recuperado el 22 de 10 de 2012, de Amazon Web Services: http://aws.amazon.com/es/elasticmapreduce/#functionality

*Apache Cassandra 0.7 Documentation*. (s.f.). Recuperado el 2012 de Agosto de 2, de Datastax: http://www.datastax.com/docs/0.7/data_model/column_families

Apache Cassandra. (14 de Junio de 2012). *Cassandra Wiki*. Recuperado el 1 de Agosto de 2012, de Apache Cassandra: http://wiki.apache.org/cassandra/

Barrios Dueñas, J. (s.f.). *Optimización de sistemas de archivo ext3 y ext4*. Recuperado el 2012 de Agosto de 6, de alcance libre: http://www.alcancelibre.org/staticpages/index.php/como-optimizar-ext3

Gartner. (s.f.). *IT Glossary*. Recuperado el 20 de 10 de 2012, de Business Intelligence (BI): http://www.gartner.com/it-glossary/business-intelligence-bi/

Gelphi Consortium. (s.f.). *The Open Graph Viz Platform*. Recuperado el 21 de 10 de 2012, de Gelphi: https://gephi.org/

Institute of Electrical and Electronics Engineers. (1990). *IEEE Standard Glossary of Software Engineering Terminology* (Vols. IEEE Std 610.12-1990). New York, NY, USA.

Jackson, J. (29 de Julio de 2011). *CouchBase, SQLite Launch Unified NoSQL Query Language.* Recuperado el 7 de Marzo de 2012, de Business Center: http://www.pcworld.com/businesscenter/article/236918/couchbase_sqlite_launch_unified_nosql_query_language.html

*JSON Tutorial*. (s.f.). Recuperado el 2012 de Agosto de 8, de w3school: http://www.w3schools.com/json/default.asp

King, R. (9 de Julio de 2010). *Engineering Blog*. Recuperado el 1 de Agosto de 2012, de Twitter: http://engineering.twitter.com/2010/07/cassandra-at-twitter-today.html

López Neira, A., & Ruiz Spohr, J. (s.f.). *ISO 27001*. Recuperado el 20 de 10 de 2012, de ¿Qué es un SGSI?: http://www.iso27000.es/sgsi.html#section2a

MSDN. (s.f.). *Escalabilidad*. Recuperado el 18 de 10 de 2012, de http://msdn.microsoft.com/es-es/library/aa292172(v=vs.71).aspx

Neo Technology. (s.f.). *System Requirement*. Recuperado el 2012 de Agosto de 4, de Neo4j: http://docs.neo4j.org/chunked/milestone/deployment-requirements.html

Neo Technology. (s.f.). *What is a Graph Database?* Recuperado el 2012 de Agosto de 4, de Neo4j: http://neo4j.org/learn/#WhatIsGraphDatabase

*NoSQL Databases*. (s.f.). Recuperado el 25 de Febrero de 2012, de NoSQL: http://nosql-database.org/

*Scalability.* (8 de Marzo de 2012). Recuperado el 12 de Marzo de 2012, de Wikipedia: http://en.wikipedia.org/wiki/Scalability

*Serial ATA Connector Pinout*. (s.f.). Recuperado el 2012 de Agosto de 6, de AllPinouts: http://www.allpinouts.org/index.php/Serial_ATA_(SATA,_Serial_Advanced_Technology_Attachment)

Shukla, I. C. (21 de Septiembre de 2011). *Different Types of Hard Drives*. Recuperado el 2012 de Agosto de 6, de Buzzle: http://www.buzzle.com/articles/different-types-of-hard-drives.html

# ANEXO 1. ENCUESTA

**Análisis de las bases de datos NoSQL como alternativa a las bases de datos SQL**

Se presenta una serie de preguntas con varias modalidades de respuesta relacionadas con la administración y uso de las bases de datos que son utilizadas por la empresa.

Las modalidades de respuesta pueden ser:

- Respuesta excluyente
- Respuesta múltiple
- Respuesta abierta

Estas preguntas representarán información valiosa para el reconocimiento de características tenidas en cuenta por las organizaciones para adquirir los diferentes motores de bases de datos.

La información recolectada por medio de esta encuesta será usada exclusivamente con fines académicos.

***Información general***:

Nombre:

Empresa:

Cargo:

1. ¿Qué bases de datos ha adquirido/usado la organización para el almacenamiento de información en los últimos 5 años?
    a. MySQL
    b. Microsoft SQL Server
    c. Oracle
    d. PostgreSQL
    e. Otro, ¿cuál/cuáles? _____

2. ¿Qué características se tuvieron en cuenta para la utilización de las bases de datos anteriores?

    a. Soporte del proveedor

    b. Integración con otras aplicaciones

    c. Facilidad de uso

    d. Velocidad en las transacciones

    e. Funcionalidad

    f. Precio

3. ¿Qué factor es más importante cuando una aplicación hace uso de la base de datos?

    a. Almacenamiento de datos.

    b. Tiempos de respuesta

4. ¿Qué funcionalidades son indispensables en un motor para que pueda hacer parte de la infraestructura de almacenamiento en la empresa?

    a. Stored Procedures

    b. Vistas (views)

    c. Triggers

    d. Funciones

    e. Event Scheduler (eventos programados)

    f. Otros, (¿cuáles?) _____

5. ¿Sobre qué sistema operativo corre el servidor de bases de datos?

    _____

6. ¿Cada cuánto tiempo realizan respaldos a la información contenida en la base de datos?

    a. Varias veces a la semana

    b. Cada semana

    c. Cada mes

    d. Lapso superior a un mes

7. ¿En alguna aplicación es necesaria la sincronización de datos en otro(s) servidor(es)?

    a. Si

    b. No

8. ¿Tienen soporte por parte del proveedor de la base de datos?

    a. Si

    b. No

9. ¿Cuánto es el costo mensual de mantenimiento y administración de una base de datos en un servidor local?

    a. Menos de $500.000

    b. Entre $500.000 y $1.000.000

    c. Más de $1.000.000

10. Cuando sale al mercado una nueva versión del motor, ¿lo adquieren inmediatamente?

    a. Si

    b. No

11. ¿Las aplicaciones actuales son en su mayoría desarrollos internos o realizados por terceros?

    a. Desarrollado en casa (interno)

    b. Desarrollado por terceros

12. ¿De forma general, ¿qué consideraciones tienen presentes para la migración de los datos a una nueva versión u otro motor de base de datos?

    _____

13. ¿Existe alguna política en la empresa que obligue a almacenar la información en instalaciones de la empresa?

    a. Si

    b. No

14. ¿Estaría dispuesto a utilizar algún sistema de almacenamiento ubicado en instalaciones de un proveedor especializado que brinde garantías de seguridad disponibilidad e integridad?

    a. Si

    b. No

15. ¿Conoce el término NoSQL?

    a. Si

    b. No

16. ¿Conoce alguna de las siguientes bases de datos?

    a. Cassandra

    b. BigTable

    c. Dynamo

    d. SimpleDB

    e. Ninguna de las anteriores

## ESCUELA DE INGENIERÍA DE ANTIOQUIA

## ACTA DE EVALUACIÓN FINAL DE TRABAJO DE GRADO

| | |
|---|---|
| **Fecha: (dd/mm/aa)** | 22/11/2012 |
| **Nombre del proyecto:** | **Análisis de las bases de datos NOSQL como alternativa a las bases de datos sql** |
| **Director del proyecto:** | Santiago Villegas Giraldo |

| **Nombre del estudiante** | **Programa académico** |
|---|---|
| **Carlos Andrés López Peña** | Ingeniería Informática |
| | |
| | |

**Nombre del Jurado:**

**Evaluación del proyecto**: Espacio exclusivo para jurado

___No aprobado __✓__Aprobado sin mención

___con Mención Pública ___ con Mención honorífica ___Trabajo laureado

**Justificación del reconocimiento**: (Artículo 28 del Acuerdo 11: "El director del Programa presentará el acta final de evaluación al Consejo Académico, donde consta la solicitud de mención especial debidamente justificada y el **Consejo determinará** si se otorga o no"). La justificación debe tener mínimo 500 palabras.

CARLOS JAIME NOREÑA MEJÍA
Director del Programa

Santiago Villegas Giraldo
Director del Trabajo de Grado

Jurado (Si lo hubo)

Jurado (Si lo hubo